BCS IRSG Symposium

Future Directions in Information Access

FDIA 2015

2nd of September 2015
Held in Thessaloniki, Greece
As part of ESSIR 2015

BCS
INFORMATION RETRIEVAL

Eds. L. Azzopardi, M. L. Wilson, I. Kompatsiaris, S. Papadaopoulos, T. Tsikrika, S. Vrochidis

2015
essir
thessaloniki.greece

SIGIR
Special Interest Group
on Information Retrieval

Elias
evaluating information access systems

YAHOO!
LABS

ALPHA BANK

BCS
INFORMATION
RETRIEVAL

Sponsors

Preface

FUTURE DIRECTIONS IN INFORMATION ACCESS

In 2007, the 1st BCS-IRSG Symposium on Future Directions in Information Access (FDIA) was established to provide a forum for early career researchers to present, share and discuss research, which is at a more formative or tentative stage. The symposium was run in conjunction with the 6th European Summer School in Information Retrieval (ESSIR) which was held in Glasgow. The second symposium was held in London, UK in September, 2008 collocated with Search Solutions, while the third symposium was again collocated with ESSIR in Padua, Italy in 2009. In 2011, it was held in Koblenz, Germany as part of ESSIR 2011 and then in Granada, Spain with ESSIR 2013. Now, in its sixth year, the 2015 Future Directions in Information Access Symposium was held as part of the 10th European Summer School in Information Retrieval.

Symposium Aims

The objectives of the Symposium on the Future Directions in Information Access are:

- To provide an accessible forum for early researchers (particularly PhD students, and researchers new to the field) to share and discuss their research.
- To create and foster formative and tentative research ideas.
- To encourage discussion and debate about new future directions.

FDIA 2015

These proceedings contain papers and posters presented at the 6th Symposium on Future Directions in Information Access, which was held in Thessalonki, Greece on the 2nd of September during the 2015 European Summer School in Information Retrieval (ESSIR).

This year's programme comprised of two parts a series of eighteen fast paced presentations followed by a poster presentation session. During the presentation phase, students gave a five-minute talk explaining their research in succinct and engaging manner, while during the poster phase presenters and participants could discuss the research in detail, form acquaintances and receive advice and mentorship from senior IR attendees. The programme featured a variety of novel and emerging topics including: temporal and location based information retrieval, visualizing user models, topic centric classification of tweets, opinionated learners, gamification of searching and learning along with work on social media analysis, emotion aware recommender systems, energy efficiency systems and reputation management systems.

The organizers would like to thank: the members of the program committee for all their hard work and effort in providing excellent feedback and reviews. We would also like to thank ESSIR 2015. We would also like to thank all the sponsors including the ACM-SIGIR for providing travel grants, the Elias Network, Yahoo! Labs, Alpha Bank and the BCS-IRSG, for sponsoring the event and the BCS EWICS Service for the online publication services.

Leif Azzopardi
Max L Wilson

September 2015

Organization

General Chair
Ioannis (Yiannis) Kompatsiaris, CERTH, Greece
Symeon Papadopoulos, CERTH, Greece
Theodora Tsikrika, CERTH, Greece
Stefanos Vrochidis, CERTH, Greece

PC Chairs
Leif Azzopardi, University of Glasgow
Max Wilson, University of Nottingham

Program Committee
M-Dyaa Albakour, University of Glasgow, UK
B. Barla Cambazoglu, Yahoo! Labs, Spain
Nicola Ferro, University of Padua, Italy
Ingo Frommholz, University of Bedfordshire, UK
Julio Gonzalo, UNED, Spain
Evangelos Kanoulas, University of Amsterdam, The Netherlands
Diane Kelly, University of North Carolina, USA
Yiannis Kompatsiaris, CERTH – ITI, Greece
Udo Kruschwitz, University of Essex, UK
Andrew Macfarlane, City University London, UK
Stefano Mizzaro, University of Udine, Italy
Michael Oakes, University of Wolverhampton, UK
Iadh Ounis, University of Glasgow, UK
Stefan Rueger, Knowledge Media Institute, UK
Tony Russell-Rose, UXLabs, UK
Fabrizio Sebastiani, Qatar Computing Research Institute, Qatar
Theodora Tsikrika, CERTH-ITI, Greece
Stefanos Vrochidis, CERTH-ITI, Greece
Leo Wanner, ICREA and University Pompeu Fabra, Spain
Max L. Wilson, University of Nottingham, UK

Table of Contents

Energy Efficiency in Web Search Engines

Matteo Catena
Gran Sasso Science Institute National Research Council of Italy
67100 L'Aquila, Italy 56124 Pisa, Italy
matteo.catena@gssi.infn.it

Today, Web search is a frequent action in the everyday life of many people. To perform it on a large scale, Web companies need energy-hungry data center, which raise environmental and economical challenges. For these reasons, Green Information Retrieval promotes energy and energy-cost awareness in contemporary Web search engines. In this document, we propose to further the research on Green Information Retrieval, which is still at its early stage. Moreover, we illustrate our first results in evaluating and improving the energy efficiency of search servers.

Web search, energy efficiency, hardware power management

1. INTRODUCTION

Web search engines continuously crawl and index large amount of web pages, which have to be promptly retrieved in response to user queries. To do so, Web search companies – e.g., Google, Yahoo!, Microsoft, Yandex, Baidu, etc. – need computer systems with large computational power and data storage capabilities. Such systems are reported to be composed by thousands of computers organized in clusters (Barroso et al. 2003), which can efficiently handle big quantities of data. These companies started building large data centers to house such computer clusters. A data center hosts large computer systems together with the associated infrastructures, such as: telecommunications, power supplying, thermal cooling, fire suppression, etc.

While data centers enable large-scale search, they also raise environmental and economical issues. The ICT sector has been reported to be responsible for roughly 2% of global carbon emissions in 2007, with general purpose data centers accounting for 14% of the ICT footprint (GeSI 2008). Moreover, power and cooling cost for 15,000 commodity servers could exceed 280,000 $/month in 2003 (Barroso et al. 2003).

For such reasons, improving data center energy efficiency has become an attractive and active research area. Nevertheless, little literature exists about energy efficiency in search engines data centers. Chowdhury is the first to explicitly write about Green Information Retrieval and to propose a research agenda for evaluating and reducing energy consumption in search services (Chowdhury 2012). In line with this agenda, we want to evaluate the energy expenditure due to the different components of a search engine. Also, we aim to investigate on possible energy saving strategies at the software (e.g., which algorithms are used to implement a component) and software architectural level (e.g., how the components are combined to form a green Information Retrieval system).

2. FIRST RESULTS

In this section, we briefly describe our initial findings on both evaluating and improving the energy efficiency of a Web search engine. At this stage, our work focuses on single search servers.

2.1. Query energy consumption

Query energy consumption is the energy consumed by a search server to solve a single query. Such information is important, since electric energy is expensive and commercial Web search engines have to keep a low cost-per-query to be profitable (Barroso et al. 2003). Moreover, recent works try to reduce search engines expenses and carbon footprint by taking into direct account their energy consumption. For instance, energy cost has been recently considered for devising energy-saving caching mechanism (Sazoglu et al. 2013). Then, precise measurements of query energy consumptions would be beneficial for such approaches.

It is possible to experimentally show that query energy consumption is linear in the query processing time. Details can be found in (Catena and Tonellotto 2015), where we experiment using the TREC ClueWeb09 corpus and MSN 2006 query log

Proceedings of the 6th Symposium on
Future Directions in Information Access 2015

to measure the energy consumption of a search server. Results reinforce the importance of efficiency improvements in Information Retrieval. More specifically, the carbon footprint of search engines can be lowered by reducing query response times without demanding additional, energy consuming, hardware. Therefore, low latencies are necessary not only to achieve user satisfaction, but also to tackle the economical and environmental costs of data centers.

2.2. Load-sensitive CPU Power Management

Typically, the energy consumption of a server is dominated by its CPU. Dynamic Frequency Scaling (DFS) technologies trade performance for reduced energy consumptions, by throttling CPU frequency (Snowdon et al. 2005). When running at low frequencies, processors absorb less power but also have lower performance than processors running at full speed. Operating systems (OS) have mechanisms that can exploit DFS to achieve energy savings. For instance, OS-level frequency governors throttle the server CPU speed accordingly to its utilization. However, the OS misses domain-specific information about the search engine application and the incoming queries.

We advocate that a more refined CPU power management is possible, knowing the search server utilization and load. In (Catena et al. 2015) we propose search engine-specific frequency governors, that manage the processor speed from within the search server application. These governors increase the CPU speed whenever the search server is struggling with processing incoming queries. Similarly, CPU speed is decreased when the search server is easily processing the arriving requests.

We conduct extensive experimentation upon the TREC ClueWeb09 corpus and the MSN 2006 query stream, to evaluate the benefits and drawbacks of our approach compared to standard OS-level frequency governors. Results show that our solution can absorb $\sim$24% less power than a system which operates at maximum CPU frequency, with only a limited detriment in query processing quality. When compared to more energy efficient OS configurations, we find that our governors can still save at least 7% in power absorption. Such energy savings are important for data centers, as reduced processor frequencies reduces heat output and thermal cooling cost. Greater energy savings can be obtained by allowing more substantial degradation in query processing quality.

3. CONCLUSIONS AND FUTURE WORK

In this work, we illustrate our first results in evaluating and improving the energy efficiency of Web search engines. Future work will continue in this direction, for instance in evaluating the energy expenditure of other search engine components (e.g., query expansion/reformulation, machine-learned document reordering, snippets generation, etc.). We also believe there is still space to further improve the CPUs power management in Web search engines data centers.

Up to this point, our work focuses on single search servers but we also wish to evaluate and improve energy efficiency at the intra-data center level, i.e., on search server clusters. Similarly, we would like to reduce energy consumption at the inter-data center level, i.e., on geographically distributed data centers owned by the same search company.

Many aspects besides hardware power management could be worth exploring. For example, one could try to understand if it is possible to trade search engines effectiveness (recall, MAP, NDCG, etc.) for energy efficiency; or what is the relationship between energy savings and corpus size; etc.

Finally, we must observe that search results are increasingly often consumed from mobile platforms (such as smartphones, tablets, etc.) with limited battery life. For this reason, future work should also promote energy efficient interactions between mobile clients and Web search engines.

4. ACKNOWLEDGMENT

The author thanks ACM SIGIR for the awarded scholarship for participating in ESSIR 2015 and the FDIA Symposium.

REFERENCES

Barroso, L.A., Dean, J., and Hoelzle, U. (2003) Web Search for a Planet: The Google Cluster Architecture. *IEEE Micro*, 23(2), 22–28.

Catena, M., Macdonald, C., and Tonellotto N. (2015) Load-sensitive CPU Power Management for Web Search Engines. To appear in: *Proc. of SIGIR*. ACM.

Catena, M., and Tonellotto, N. (2015) A Study on Query Energy Consumption in Web Search Engines. In: *Proc. of IIR*. CEUR-WS, `http://ceur-ws.org/Vol-1404/`.

Chowdhury, G. (2012) An agenda for green information retrieval research. *Information Processing & Management*, 48(6), 1067–1077.

Sazoglu, F.B., Cambazoglu, B.B., Ozcan, R., Altingovde, I.S., and Ulusoy, O. (2013) A Financial Cost Metric for Result Caching. In: *Proc. of SIGIR*. ACM, 873–876.

Snowdon, D.C, Ruocco, S., and Heiser, G. (2005) Power Management and Dynamic Voltage Scaling: Myths and Facts. In: *Proc. of PARC Workshop*.

The Climate Group for GeSI (2008) *Smart 2020: Enabling the low carbon economy in the information age*. Available from: `http://goo.gl/4I3vun` (16 June 2015).

A Gamification Framework for Enhancing Search Literacy

Ioannis Karatassis
Department of Computer Science
and Applied Cognitive Science
University of Duisburg-Essen, Germany
karatassis@is.inf.uni-due.de

Recent studies reveal that the overall search literacy leaves something to be desired and that the most people overestimate their skills in the domain of Web search. In this paper, a gamification framework is introduced that aims at increasing the search literacy within a ludic environment to support Internet users in being more successful in their daily search sessions. Furthermore, the paper presents plans for future work that strive for answering whether the search literacy can be improved by applying the presented approach and how we can detect and especially measure such an improvement.

search literacy, gamification, Web search, enhancement

1. INTRODUCTION

As the size of information on the World Wide Web increases, the effective use of Web search engines for information retrieval (IR) becomes a key challenge for Internet users. Search engines act as gatekeepers and provide access to online resources. Internet users are facing the challenge of finding the desired information and need therefore skills that can be broadly summarized under the term “information literacy”. A person is information literate if she is able to recognize when information is needed and has the ability to locate, evaluate, and use effectively the needed information. This paper focuses on a specific aspect of information literacy – the search literacy. The latter concept refers directly to the retrieval process of information and denotes the abilities to locate and access desired information in order to satisfy information needs with efficiency and effectiveness.

A recent study revealed that most users have a rather limited search literacy as well as they overestimate their competence in this area (Stark et al. (2014)). Bateman et al. (2012) stated that most searchers do not know how to use search engines effectively for satisfying their information needs. One possible reason is that beyond query completion, search engines provide no feedback to users that would help them in improving their search behavior. White (2013) observed that users struggle with finding an answer to a yes-no question; moreover, biases arising from people’s beliefs influence their judgment, decision making and actions. Kodagoda and Wong (2008) showed that low literacy users take significantly more time than high literacy users to complete an information task, and were significantly less accurate. In addition, low literacy participants spent more time on a Web page to find the desired information.

These findings show that there is a need to improve the search literacy of Web search engine users. In this paper, we describe an approach for enhancing search skills via gamification in the domain of Web search engines to help users being more successful in their searches. For this purpose, we have developed a gamification framework that features different kinds of tasks including search and educational tasks in terms of Web search engines. The main goal is the gamification of the core IR task, i.e., Web search. We aspire noticeable and sustainable improvements of skills in the context of Web-based IR.

2. RELATED WORK

2.1. Search literacy

Producing accurate search results requires knowledge about the basic functioning of search engines as well as the following aspects: searchability, linguistic functions, query language, and ranking (Fuhr (2014)). According to Fuhr, a search-literate user needs to know appropriate search tactics and strategies in order to succeed in satisfying information needs effectively.

Searchability Users need to be aware of unsearchable content, since not all online resources that can be accessed via browsers, are findable. For example, the language used in the search query, the document

© *The Authors. Published by BCS
Learning and Development Ltd.
Proceedings of the 6th Symposium on
Future Directions in Information Access 2015*

type, and the recency of Web pages are some reasons why Web resources can not be found.

Linguistic functions Search engines apply linguistic functions to search queries such as word normalization, lemmatization, and phrase identification, and take composites and synonyms into account to deal both with the vagueness and the ambiguity of natural languages which form a crucial IR problem.

Query language The use of search operators (e.g., Boolean operators, number ranges, facets, and field and URL predicates) and search options (e.g., for restricting the time, place, language, and document type of result items) allows users to express complex information needs and leads to more specific search results.

Ranking Studies exposed that users prefer search results on the first result page; items below the fold are seldom clicked on (Höchstötter and Lewandowski (2009)). Hence, it is important to produce result sets where relevant documents are located at a top position on the first page and in the visible area without the need to scroll by formulating precise queries for not missing relevant information.

Tactics and strategies Expressing complex information needs usually requires a series of steps and search queries. Strategies are plans for performing a complex search whereas tactics denote single operations to advance searches. Bates (1979) distinguishes between the following types of tactics: monitoring, file structure, search formulation, and term.

2.2. Gamification

Gamification is defined as a process of integrating game mechanics into non-game contexts to invoke gameful experiences and to engage users in solving problems (Zichermann and Cunningham (2011)). On the other side, if applied improperly, gamification can reduce the internal motivation or even replace it by external motivation (Nicholson (2012)). The concept is a viable means for increasing users' motivation, shaping users' behavior, and enhancing online services with (motivational) affordances to invoke ludic experiences. It primarily takes advantage of the fact that games are fun (Poels et al. (2007)). When developing gamification strategies, designers have to be aware of different kinds of gamers since each player has different motivations, in-game behaviors, and play styles (Dixon (2011)). Although there is still a lack of empirical evidence on the side effects of employable game elements, the findings of various studies (e.g., Hakulinen et al. (2013); O'Donovan et al. (2013)) lead to the conclusion that gamification does not harm the internal motivation of users at all if applied in a user-centered fashion (Nicholson (2012)). On the contrary, points, levels, leaderboards, and badges are an easy and effective way to increase users' performance. Nevertheless, one should take social and contextual factors into account as they may determine whether the employed game elements diminish (see Mekler et al. (2013)) or even suppress internal motivation.

3. FRAMEWORK

Traditional training methods and tutorials are meant to be a potential means to enhance the abilities of people. We focus on gamification where users have to master problems in a playful manner. Our approach aims at increasing users' Web search skills and basic understanding of the functioning of Web search engines. By "playing our game", users are to learn and to develop sustainable Web search strategies for satisfying their information needs subconsciously. In the best case, people will be aware of their new skills. We expect users noticing a (significant) improvement of their Web search skills as well as the precision of the search results.

3.1. Architecture

Our Web framework consists of a client-side (front end) and server-side (back end) component where appropriate handlers are capable of communicating with each other to handle user input and to change the application state as a result. Front end handlers validate and forward user input, received from user interfaces, via messages to the corresponding back end handlers that are interconnected with other modules. We put emphasis on a module based architecture that allows for an easy extension by features, tasks, and game design elements. In this context, a database serves for storing task-, user- and application-related data. The front end is accessible via modern browsers and features a high usability (see section 4) that allows users to achieve the specified goals with effectiveness, efficiency and satisfaction.

The core of our developed framework is a search proxy that is used to retrieve search results from Web search engines. Our main goal is the gamification of Web search where users have to solve interactive search tasks. Microsoft's Bing[1] suits our needs, since it offers a wide range of features. We exploit this rich feature set to provide a complete search interface containing all commonly employed search functions to our users. Depending on the received search query and search options from back end handlers, the proxy creates a URL to access the Bing search API. Search results are received in textual form, i.e., in JSON format, and need therefore to be parsed first. Once appropriate objects have been created, the results are forwarded to the initial caller, the front end handler, which renders them along other information according to task related settings.

Basically, users are provided with different types of tasks they have to solve that allow the application of various strategies to enhance search literacy on different levels.

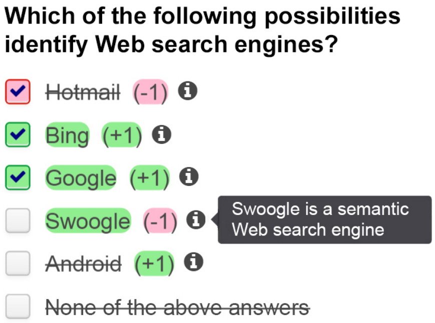
[1] http://www.bing.com/

To emphasize the playful character of the application, we introduced the notion of “game mode” – each game mode is responsible for one type of task. Being able to detect an improvement of search skills requires keeping track of a user’s progress in the “game”. In this context, the logging of user interactions and corresponding results contributes to the creation of user profiles in order to analyze and to disclose the “play behavior” of our users.

3.2. Game modes

3.2.1. Quiz

This game mode aims at determining as well as increasing the basic knowledge of users related to Web search engines. Users are presented with questions they have to answer in terms of selection. Depending on the question type (single or multiple choice), one has to select preset answers one considers to be correct.



Which of the following possibilities identify Web search engines?

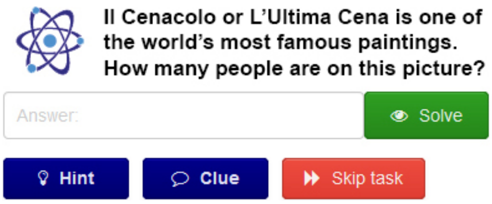
☑ ~~Hotmail~~ (-1)
☑ Bing (+1)
☑ Google (+1)
□ Swoogle (-1)
□ ~~Android~~ (+1)
□ ~~None of the above answers~~

Swoogle is a semantic Web search engine

Figure 1: *Additional information that contributes to the understanding of the answer can be accessed through the information icon next to each one.*

3.2.2. Search hunt

The purpose of this game mode is to support users in increasing the precision of their search queries and in improving their relevance judgment and content-finding skills. In search hunt tasks, we ask questions for which users have to find the corresponding answers using a feature-rich search interface that is connected to the aforementioned search proxy. Answers are to be found by formulating search queries and identifying as well as exploring relevant search results.

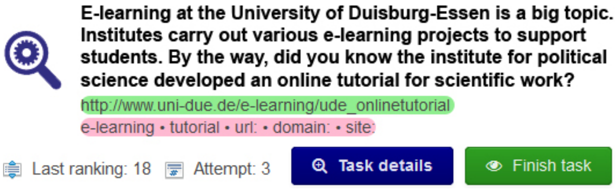


Il Cenacolo or L’Ultima Cena is one of the world’s most famous paintings. How many people are on this picture?

Figure 2: *In this scenario, a user has to create a search query to find the desired image. The task interface offers hints and clues that users can request for a fee, i.e., points, as well as an option for skipping the current task without evaluation if desired.*

3.2.3. Query tuning

As the name suggests, the game mode aims directly at increasing the precision of search queries. However, the methods differ from search hunts. A Web page (URL) and related details are supplied. Users have to create a search query that ranks the given Web page at a top position within the results list received by the search proxy. The closer to the first position the given Web page is, the better users are rewarded. While refining (tuning) search queries to solve the task, searchers are also to reflect on the changes occurring in the results where even small modifications to the query may lead to very different results.



E-learning at the University of Duisburg-Essen is a big topic. Institutes carry out various e-learning projects to support students. By the way, did you know the institute for political science developed an online tutorial for scientific work?

http://www.uni-due.de/e-learning/ude_onlinetutorial

e-learning • tutorial • url: • domain: • site:

Last ranking: 18 Attempt: 3

Figure 3: *Users have to enter a search query in a text field which will be forwarded via the search proxy to the connected search engine to rank the URL marked in green at a top position. The words marked in red are terms or search operators that must not appear in the search query. Users are free to finish a task at any time. The evaluation encompasses the last search process and the amount of needed attempts.*

3.3. Game design elements

The framework is provided with points, levels, badges, leaderboards, and sound effects. Points are received for solving tasks partially or fully correct. In addition, one can receive bonus points, e.g., for quick solving tasks. Levels specify a user’s current game state and the task difficulty: the higher the level is, the more difficult are the tasks users are presented with. By exceeding point thresholds, one can reach the next levels. User interactions such as “solving a task” or “reaching x points in game mode y” are connected with badges. When triggered, the user is rewarded accordingly in case the appropriate conditions are met. Leaderboards – one per game mode – allow simple comparisons, promote competition, and serve as a powerful motivator for continuing. We represent leaderboards as ordered lists with a point score beside each name. Along with levels, leaderboards indicate that players have status or achievement in the game. Sound effects are used to guide users and to introduce events, such as the receipt of points/badges, the beginning/ending of a task, and the reaching of the next level.

4. FIRST RESULTS

First of all, we measured the usability to test whether the user interface is usable. Concerning this matter, we created demonstration tasks for each game mode that have been solved during a user test by $N = 15$ participants. At the end of each lab session, the System Usability Scale (SUS) was used for measuring perceptions of usability. A total score of 90.2 allows us to conclude that the framework (or more precisely, the user interface) is user-friendly and capable of increasing users’ motivation.

5. CONCLUSION AND OUTLOOK

We created a framework that features a high usability in order to fulfill the desired goal effectively and expect to improve the quality of search sessions which will lead in general to a higher task completion rate. The next step is to create a rich set of challenging tasks to train Web search engine users. Each game mode can be used to enhance different aspects of search literacy. For instance, quiz tasks provide more insights into the functioning of search engines. Search hunt tasks can be designed in a manner that the users have to explore various commonly employed search engine interface features, such as search options or operators, to find appropriate answers. As different result types provide different options, solutions can be located on Web results as well as on images, news, or even in videos. Query tuning tasks support users in creating precise search queries within a step-by-step refining process.

After the creation of our task set, we will run a long-term study with a large user base. The outcomes of this study will help us to tune our system and to generate ideal solutions for each task which we will use to score individual users by the closeness of their solution to the ideal one. Furthermore, we will identify key factors that make a user *search literate* and compare users' knowledge to the initial one to answer whether and under what circumstances the search literacy can actually be improved. In addition, we plan to invite once more interested participants after a certain period of time who will solve new tasks but with the same complexity in order to test the sustainability of the improvements.

REFERENCES

S. Bateman, J. Teevan, and R. W. White. The search dashboard: How reflection and comparison impact search behavior. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 1785–1794, New York, NY, USA, 2012. ACM.

M. J. Bates. Information search tactics. *Journal of the American Society for Information Science*, 30(4): 205–214, 1979.

D. Dixon. Player types and gamification. *Proceedings of the CHI 2011 Workshop on Gamification*, 2011.

N. Fuhr. Internet search engines – Lecture script for the course in SS 2014, 2014. Available online at http://www.is.inf.uni-due.de/courses/ir_ss14/ISMs_1-7.pdf (in German); accessed 16-June-2015.

L. Hakulinen, T. Auvinen, and A. Korhonen. Empirical study on the effect of achievement badges in trakla2 online learning environment. In *Proceedings of the 2013 Learning and Teaching in Computing and Engineering*, LATICE '13, pages 47–54, Washington, DC, USA, 2013. IEEE Computer Society.

N. Höchstötter and D. Lewandowski. What users see - structures in search engine results pages. *Inf. Sci.*, 179(12):1796–1812, May 2009. ISSN 0020-0255.

N. Kodagoda and B. L. W. Wong. Effects of low & high literacy on user performance in information search and retrieval. In *Proceedings of the 22Nd British HCI Group Annual Conference on People and Computers: Culture, Creativity, Interaction - Volume 1*, BCS-HCI '08, pages 173–181, Swinton, UK, 2008. British Computer Society.

E. D. Mekler, F. Brühlmann, K. Opwis, and A. N. Tuch. Do points, levels and leaderboards harm intrinsic motivation?: An empirical analysis of common gamification elements. In *Proceedings of the First International Conference on Gameful Design, Research, and Applications*, Gamification '13, pages 66–73, New York, NY, USA, 2013. ACM.

S. Nicholson. A User–Centered Theoretical Framework for Meaningful Gamification. Paper presented at Games+Learning+Society 8.0, Madison, WI, June 2012.

S. O'Donovan, J. Gain, and P. Marais. A case study in the gamification of a university-level games development course. In *Proceedings of the South African Institute for Computer Scientists and Information Technologists Conference*, SAICSIT '13, pages 242–251, New York, NY, USA, 2013. ACM.

K. Poels, Y. de Kort, and W. Ijsselsteijn. "It is Always a Lot of Fun!": Exploring Dimensions of Digital Game Experience Using Focus Group Methodology. In *Proceedings of the 2007 Conference on Future Play*, Future Play '07, pages 83–89, New York, NY, USA, 2007. ACM.

B. Stark, D. Dörr, and S. Aufenanger. The Google-ization of information search – Search engines in the field of tension between usage and regulation. Management Summary, 2014. Available online at http://www.ifp.uni-mainz.de/Bilder_allgemein/Suchmaschinen_Management_Summary.pdf (in German); accessed 16-June-2015.

R. White. Beliefs and biases in web search. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 3–12, New York, NY, USA, 2013. ACM.

G. Zichermann and C. Cunningham. *Gamification by Design: Implementing Game Mechanics in Web and Mobile Apps*. O'Reilly Media, Inc., 1st edition, 2011.

A Framework for Emotion-aware Recommender Systems supporting Decision Making

Marco Polignano
Ph.D. Student at University of Bari Aldo Moro
Via Orabona 4, Bari, Italy
www.di.uniba.it/ swap
marco.polignano@uniba.it

Emotions influence everyday decisions. When people make decisions about movies to watch, songs to listen or even about more serious issues such as health, they perform a cognitive process that estimates which of various alternative choices would yield the most positive consequences. Indeed, this process in not totally rational because it is influenced, directly or in a subtle way by personality traits and emotions. In this paper we propose the idea of defining an affective user profile, which can act as a computational model of personality and emotions, included in a general, affective-aware, recommendation framework.

Emotions, Personality traits, Recommender Systems, Human Decision Making

1. BACKGROUND AND MOTIVATION

The question of how to conceptualize emotions concerning their role in decision making (DM) has been deeply studied in the psychological literature over the last twenty years Pfister (1992, 2008); Loewenstein et al. (2003); Peters (2006); Fiori (2013). According to traditional approaches of behavioural decision making, choosing is seen as a rational cognitive process that estimates which of various alternative choices would yield the most positive consequences, which does not necessarily entail emotions. Emotions are considered as external forces influencing an otherwise non-emotional process (influence-on metaphor). Loewenstein et al. (2003) distinguish between two different ways in which emotions enter into decision making. The first influence is that of expected emotions, i.e. beliefs about the emotional consequences of the decision outcomes. Users might evaluate the consequences of the possible options by taking into account both positive and negative emotions associated with them and then select those actions that maximize positive emotions and minimize negative emotions. The other kind of affective influence on DM consists of immediate emotions that are experienced at the time of decision making. Such feelings often drive behaviour in directions that are different from those coming from the rational mental process and thus derived by a consequentiality evaluation of future consequences. Recommender Systems are tools that support users in the process of choosing options in several domains. These systems, usually implementing an information filtering algorithm based on user preferences, are starting to consider also affective information to adapt suggestions according to "emotional features" of users or items.

The positive influence of affects in recommender systems is shown by Zheng and Burke (2013) that got relevant results in terms of increased precision in Context-Aware Recommender Systems. Similar results are obtained by Tkalcic et al. (2013), that show an increase of performance in content-based recommender systems that use emotional item labelling.

In recommender systems literature, emotional feedback play different roles related to the acquisition of user preferences:

1. As a source of affective meta-data for item modelling and building a preference model;
2. As an implicit relevance feedback for assessing user satisfaction.

In this work, we focus on the first issue: the idea is to acquire affective features that might be

Proceedings of the 6th Symposium on Future Directions in Information Access 2015

exploited for user modelling. The aim is to define a general framework to include emotional aspects into a user profile augmented with affective features that can be exploited in the process of computing recommendations. In particular, we would like to design a novel, general recommendation process that takes into account both user personality traits and immediate emotions. For this purpose, three tasks must be performed:

1. Emotions identification;
2. Emotions formalization into the affective user profile;
3. Design of a recommendation process based on the affective user profile.

The work is still at an early stage and preliminary ideas will be discussed.

2. STRATEGIES TO IDENTIFY EMOTIONS

In Tkalcic et al. (2011), the authors show how it is possible to identify emotions during three different steps of the interaction between users and recommender systems. Following this structuring of user-system interaction, we will detect emotions in these different steps, using specific acquisition strategies:

1. Early stage: in this stage the user will face the decision task coming from an external context. The emotions which the user feels facing the problem will be detected. Those feelings are consequence of events not correlated to the decision, thus they should be found outside the recommendation process. For instance, they could be gathered from user's account on social networks like Facebook or Twitter. Daily posts will be analysed using sentiment analysis techniques to identify the user's affective state (e.g. mood) in the early stage of the decision task. Furthermore, A Big Five Inventory questionnaire proposed by John and Srivastava (1999) will be used to get the user personality traits (this could be done only once).

2. Consumption stage: during the decision task, the user will face with their expectation and their expected emotions after the decision. The current affective state could be acquired through explicit questions that allow the categorization of the current emotion among the six Ekman universal emotions Ekman (1993): happiness, sadness, surprise, fear, disgust, anger. Implicit affective feedback could be acquired by monitoring the user facial expression through specific tools for emotion detection.

3. Exit Stage: at the end of the decision task, consequence emotions will be available. Consequences could be immediate or postdated. If decision consequences are immediate, exit stage emotions will be gathered using the same techniques adopted during the consumption stage. An analysis of user's social network posts could be also performed to identify consequence emotions.

   The uthor identify two different main category of methods for gather emotions information: the explicit ask and implicit detection.

3. AFFECTIVE PROFILE

The user's emotions will be stored as components of her affective profile. Every decision taken using the system, will be stored formally to become an affective historical case for the specific domain. The history of the user decisions is the primary base of knowledge for a recommender system that must support the decision making task. This affective-augmented knowledge base allows to compute the user preferences with respect to an option, in a specific context, while it is affected by a well defined emotion. The affective profile is composed by user personality traits (PT), historical decision cases (HC), contexts and user expertise (CE).

$$AP = PT \times HC \times CE \tag{1}$$

Personality Traits. Personality traits are formalized as a distribution of percentage values among the dimensions: Openness to experience, Conscientiousness, Extroversion, Agreeableness, Neuroticism in according to the Big Five model described by Goldberg (1990). These elements are the distinctive traits of the user behaviour which allow to predict user common preferences and decisions. A demonstration of these theories in a social network context is provided by Moore and McElroy (2012).

Historical decision cases. The historical case, stored in the user profile, is a formalization of the decision taken. It must describe accurately the decision task and emotions felt by the users.
The emotional state will be formalized as a distribution of the six Ekman emotions during the decision process. The task is defined by the context of decision, the faced problem, options to choose, decision taken, feedback in a scale from 1 to 10 to describe the utility of suggestions (1 not useful, 10 extremely useful).

Context and expertise. In these segment of the affect profile, all the contexts faced from the user are stored. For each context will be stored also the specific user preferences and skills to allow the recommender system to better understand the user's needs. We can formalize the expertise of each skill associating it with the number of decisions taken in this context.

4. EMOTION-AWARE RECOMMENDATION PROCESS

Recommender Systems (RSs) are largely used in a lot of different domains, from the classical e-commerce system to the more risky financial advisory domain. Commonly they are based on user's or item's descriptive features but they do not consider users irrational features such as emotions. An Emotion-Aware Recommender System takes as input information from the user affective profile and generates solutions to support the user in the decision task taking into account emotionally attributes. The recommendation strategy is base on the Case-based reasoning one of the most commonly adopted machine learning method, that exploits a knowledge-based representation of the context.

An Emotion-Aware Recommender System has, first, to use similarity measures to identify users that match the active user's affective profile. In particular they are used on vectors that includes user's personality features and preferences in the specific context. For each user identified, including the active user, are gathered historical cases that match the problem, the active user early stage emotional state, and positive exit stage emotions or positive user feedback. From the historical cases detected, candidate solutions are extracted and filtered or ranked according to the context of the problem. Using a preliminary week classification based on the level of risk of the decisions,three different macro contexts could be identified for our framework:

1. High risk domains: hard decisions are taken. In this context, it is important to provide appropriate and understandable (i.e. explainable) solutions. The important aspect to be taken into account is the correctness. An application that falls in this category is a recommender systems for financial investments. In this case, the RS could decide to mitigate the negative emotions felt by an inexperienced user by proposing low-risk investments.

2. Medium risk domains: decisions in these domains can hardly be reversed. An application that fall in this category is a RS for activity plans. In this context, there are some constraints that have to be satisfied, for example, work commitments.

3. Low risk domains: decisions in these domains are easy to revert. It is possible for the RS to suggest new and uncommon items, by diversifying recommendations according to preferences and emotional state of the user. An application that fall in this category is a music recommender system which can propose play-list according to the user mood and her tendency to maintain or change it based on her personality traits.

The framework will be designed according to the described recommendation approach and macro-categoritazion of domains for the ranking of the solutions. When poor historical data are avaible, the described pipeline is not efficient. To supply the problem will be used strategies of inference of preferences from user's personality traits. For example, happy items will be suggested to users who have hight agreeableness value. An empirical demonstration of correlations between personality and user's preferences in a music domain is provided by Ferwerda (2015).

5. CONCLUDING REMARKS AND ONGOING WORK

Emotions are important elements of people's life. In each decision making task, emotions influence the choosing process. In those contexts in which decisions lead to risky consequences, emotions need to be mitigated, while in others, such as music recommendation, they could be amplified and used to generate useful suggestions. Systems that support the decision making task, currently take into account emotions in a limited way, while we have proposed a framework able to embed emotions and personality traits into the recommendation process. The ideas proposed in this paper are currently developed within the doctoral program of the author, therefore they are still at a preliminary stage.

6. ACKNOWLEDGEMENTS

Thanks to ACM Special Interest Group on Information Retrieval (ACM SIGIR) to provides a scholarship for attend the Symposium on Future Directions in Information Access (FDIA) 2015 and the 10th European Summer School in Information Retrieval. Thanks to Marco de Gemmis ”Dept. of Computer Science, University of Bari Aldo Moro” supervisor of the work.

REFERENCES

Gerald L. Clore (1992) Cognitive Phenomenology: Feelings and the Construction of Judgment. In L. Martin. Tesser (eds.), The Construction of Social Judgments. Lawrence Erlbaum. Pages: 10-133

Charles R. Darwin (1872) The expression of the emotions in man and animals. London: John Murray. 1st edition

Paul Ekman (1993) Facial expression and emotion. American Psychologist

Bruce Ferwerda, Markus Schedl, Marko Tkalcic (2015) Personality & Emotional States: Understanding Users Music Listening Needs. In A. Cristea, J. Masthoff, A. Said, & N. Tintarev (Eds.), UMAP 2015 Extended Proceedings

Marina Fiori, A. Lintas, S. Mesrobian, and A. E. P. Villa (2013) Effect of emotion and personality on deviation from purely rational decision-making. Decision Making and Imperfection, volume 474 of Studies in Computational Intelligence, Springer. Pages: 129-161

Lewis R. Goldberg (1990) An alternative description of personality: the big-five factor structure. In: Journal of personality and social psychology 59.6. Pages: 1216

Oliver P. John, Sanjay Srivastava (1999) The Big Five trait taxonomy: History, measurement, and theoretical perspectives. In: Handbook of personality: Theory and research 2, 1999. Pages: 102-138

George F. Loewenstein, Elke U. Weber, Christopher K. Hsee, Ned Welch (2001) Risk as feelings. In: Psychological bulletin, 127(2), From page: 267

George F. Loewenstein and J. S. Lerner (2003) The role of affect in decision making. Oxford University Press. Pages 619-642

Kelly Moore and James C. McElroy (2012) The influence of personality on Facebook usage, wall postings, and regret. In: Computers in Human Behavior 28.1, 2012. Pages: 267-274

Ellen Peters (2006) The functions of affect in the construction of preferences. In: The construction of preference. Pages: 454-463

Hans-Rdiger Pfister and G. Böhm (1992) The function of concrete emotions in rational decision making. Acta Psychologica 80. Pages: 199-211

Hans-Rudiger Pfister and G. Bohm (2008) The multiplicity of emotions: a framework of emotional functions in decision-making. Judgment and decision making 3(1). Pages: 5-17

Schlosser, Thomas, David Dunning, and Detlef Fetchenhauer (2001) What a feeling: the role of immediate and anticipated emotions in risky decisions. In: Journal of Behavioral Decision Making 26.1. Pages: 13-30

Marko Tkalcic, Andrej Kosir and Jurij Tasic (2011) Affective recommender systems: the role of emotions in recommender systems. RecSys11 Workshop on Human Decision Making in Recommender Systems In conjunction with the 5th ACM Conference on Recommender Systems, Chicago, IL, USA, October 23-27

Marko Tkalcic, Ante Odic, Andrej Kosir, Jurij Tasic (2013) Affective Labeling in a Content-Based Recommender System for Images. In: IEEE transactions on multimedia, vol. 15, no. 2, February. Pages: 391-400

Yong Zheng, Bamshad Mobasher, and Robin D. Burke (2013) The Role of Emotions in Context-aware Recommendation. In: Decisions@ RecSys, 2013. Pages: 21-28

Improving Information Retrieval Evaluation via Markovian User Models and Visual Analytics

Maria Maistro
Department of Information Engineering
University of Padua, Italy
maistro@dei.unipd.it

To address the challenge of adapting experimental evaluation to the constantly evolving user tasks and needs, we develop a new family of Markovian Information Retrieval (IR) evaluation measures, called Markov Precision (MP), where the interaction between the user and the ranked result list is modelled via Markov chains, and which will be able to explicitly link lab-style and on-line evaluation methods. Moreover, since experimental results are often not so easy to understand, we will develop a Web-based Visual Analytics (VA) prototype where an animated state diagram of the Markov chain will explain how the user is interacting with the ranked result list in order to offer a support for a careful failure analysis.

Evaluation, Markov Precision, User Model, Visual Analytics

1. INTRODUCTION

Nowadays information and its retrieval are fundamental and pervasive in everyday life of each person and the methods of accessing it and the systems themselves are changing rapidly. Hence, the quantity and heterogeneity of available information is rapidly increasing as well as the complexity of user tasks and needs are performing. This calls for increasingly sophisticated IR methods and systems which, in turn, need advanced evaluation techniques to be properly conceived, designed and developed.

In particular, IR systems operate using a best match approach: in response to an often vague user query, they return a ranked list of documents ordered by the estimation of their relevance to that query. In this context effectiveness, meant as “the ability of the system to retrieve relevant documents while at the same time suppressing the retrieval of non-relevant documents” (Rijsbergen 1979), is the primary concern. Since there are no a-priori exact answers to a user query, experimental evaluation based on effectiveness is the main driver of research and innovation in the field. Indeed, the measurement of system performances from the effectiveness point of view is basically the only mean to determine which are the best approaches and to understand how to improve IR systems.

Today the available evaluation methods can be divided in two distinct categories: the batch and the on-line methods. The batch methods are based on models that consider a hypothetical user and depict his/her behaviour in an abstract way. Hence, the user does not interact with the system and the system is evaluated through controlled experiments. The main disadvantage of these methodologies is that they are mostly focused on the algorithmic and system side and they may be somewhat “artificial” by abstracting away too much of the user interaction. On the other hand, these methods have the advantage of being reproducible and scalable.

The other family of evaluation methods, the on-line methods, bases its strategy on user studies and analysis of interaction data, such as search logs, to investigate and consequently connect the behaviour of the user with the system. These methods undoubtedly have the advantage of taking into account the user’s needs and to interact with him/her, but they are more expensive because they require the involvement of real users, they are time consuming and they cover many different disciplines. Furthermore on-line methods are not easily reproducible or scalable.

In batch evaluation, many different measures, e.g. precision and recall, have been created to determine in a rigorous way when one ranked list of documents is better sorted than another one. However, if the value of precision or the

Proceedings of the 6th Symposium on
Future Directions in Information Access 2015.

value of recall increase, does user satisfaction also increase? It is therefore fundamental to envision new evaluation methodologies capable of linking on-line and batch strategies and of providing a better fit with actual user needs and behaviour. This is needed to provide a more accurate estimation of system performances, which is crucial to cope with ever increasing information resources and rapidly evolving user tasks.

Moreover, both batch and on-line evaluation methods produce huge amounts of experimental and scientific data from which it is not so obvious how to infer useful information. Statistical tools (Savoy 1997) and other recent techniques, for instance VA (Keim et al. 2010), can play a key role in coping with and understanding such large amounts of experimental data. Thus, new evaluation methodologies should also comprise powerful VA techniques to support researchers and developers in analysing, exploring and understanding experimental results in order to more effectively improve IR systems.

2. OBJECTIVES

As previously discussed, there are two methods that allow for the evaluation of an IR system: batch and on-line methods. An evaluation metric including both these aspects represents an innovation with respect to the state-of-the-art because:

- up to now there is not a measure capable of connecting the lab-style and the on-line evaluation methods and to merge them in a single tool, also accounting for the time dimension;
- from the scientific point of view this will provide the possibility of fostering an in-depth analysis of user behaviour models;
- From the engineering point of view, a more powerful measure, able to better grasp and explain the interaction between the system and its users, will provide a valuable support for the design and the development of next generation IR systems.

To reach the purpose of defining a class of measures which can be used with both batch and on-line strategies, we plan to rely on the Markov chains framework (Norris 1998) as proposed in our work (Ferrante et al. 2014). Furthermore, regarding the management and the visualisation of the experimental data we will use VA techniques, which are a quite new idea to the IR field (Angelini et al. 2014), and which allow the experimental results to be more efficiently and effectively explained.

Therefore, we can summarize the aim of this research project in four main objectives:

Definition of the Markov measures : a new family of metrics based on Markov chains that can be used both with the batch and the on-line evaluation methods;

Analysis of the properties of the Markov measures both from the mathematical and experimental point of view;

Design and development of a prototype web application that uses techniques of VA to represent experimental results;

Evaluation of the web prototype with domain experts and examples of its possible applications.

3. METHODOLOGIES

In this Section we will describe the methodologies adopted to reach the four principal objectives and the first research results.

3.1. Definition of the Markov Measures

Even though Markov chains are an intuitive and robust tool, up to now they have never been applied to the field of evaluation in IR, except for our paper (Ferrante et al. 2014). In particular, we define a new class of evaluation measures called MP, by representing each position in a ranked result list with a state in a Markov chain and the transition probabilities among the states allow us to model the different and complex user interaction in scanning the ranked result list, e.g. forward and backward movements or jumps.

Firstly we introduce some notation that we use through this section. Let us consider a ranked list of T documents, let $\mathcal{R}$ be the set of the ranks of the relevant documents and RB the recall base, i.e. the total number of judged relevant documents. We assume that each user starts from a chosen document, at rank X_0 in the list, and considers this document for a random time T_0, that is distributed according to a known positive random variable. Then he/she decides to move to another document, at rank X_1, and he/she considers this new document for a random time T_1. Successively, he/she moves, independently, to a third document and so on. Hence, we denote by $X_0, X_1, X_2, \dots$ the (random) sequence of document ranks visited by the user and by T_0, T_1, T_2 the random times spent visiting each considered document.

We mathematically model the user behaviour in the framework of the Markovian processes by assuming that X_0 is a random variable on $\mathcal{T} = \{1, 2, \dots, T\}$

with a given distribution $\lambda = (\lambda_1, \dots, \lambda_T)$; so for any $i \in \mathcal{T}$, $\mathbb{P}[X_0 = i] = \lambda_i$. Then, we assume that the probability to pass from the document at rank i to the document at rank j will only depend on the starting rank i and not on the whole list of documents visited before. Thanks to this condition and fixing a starting distribution λ, the random variables $(X_n)_{n \in \mathbb{N}}$ define a time homogeneous discrete time Markov Chain, with state space $\mathcal{T}$, initial distribution λ and transition matrix P.

To obtain a continuous-time Markov Chain, we have to assume that the holding times T_n have all exponential distribution and conditioned on the fact that $X_n = i$, the law of T_n will be exponential with parameter μ_i, where $\mu_i > 0$. When our interest is only on the jump chain $(X_n)_{n \in \mathbb{N}}$, we simply assume that all these variables are exponential with parameter $\mu = 1$; while when we are also interested in the time dimension, we have to provide a calibration for these exponential variables.

Let us assume hereafter that the matrix P is irreducible and that after visiting n documents in the list the user will stop his/her search. In order to measure his/her satisfaction, we evaluate the average of the precision, computed at the ranks of the relevant documents visited by the user, as

$$\frac{1}{n} \sum_{k=0}^{n-1} \text{Prec}(Y_k) \ ,$$

where $(Y_n)_{n \in \mathbb{N}}$ denotes the sub-chain of $(X_n)_{n \in \mathbb{N}}$ that considers just the visits to the judged relevant documents at ranks $\mathcal{R}$. Clearly, this quantity is of little use if evaluated at an unknown finite step n. However, the Ergodic Theorem for the Markov processes approximates this quantity with

$$MP = \sum_{i \in \mathcal{R}} \pi_i \text{Prec}(i) \quad ,$$

where π is the (unique) invariant distribution of the Markov chain $(Y_n)_{n \in \mathbb{N}}$. Note that MP is defined without knowing the recall base RB, but just the ranks of the judged relevant documents in the given run.

In order to include the time dimension, we can replicate the previous computations and define a new measure

$$MPcont = \sum_{i \in \mathcal{R}} \widetilde{\pi}_i \text{Prec}(i).$$

where $\widetilde{\pi}_i = \frac{\pi_i (\mu_i)^{-1}}{\sum_{j \in \mathcal{R}} \pi_j (\mu_j)^{-1}}$, π denotes again the (unique) invariant distribution of the Markov chain $(Y_n)_{n \in \mathbb{N}}$ and μ_i is the parameter of the holding time in state i.

Therefore, when we consider the discrete-time Markov chain, we are basically reasoning as traditional evaluation measures which assess the utility for the user in scanning the ranked result list (batch measure), while when we consider the continuous-time Markov chain, we also embed the information about the time spent by the user in visiting a document(on-line measure).

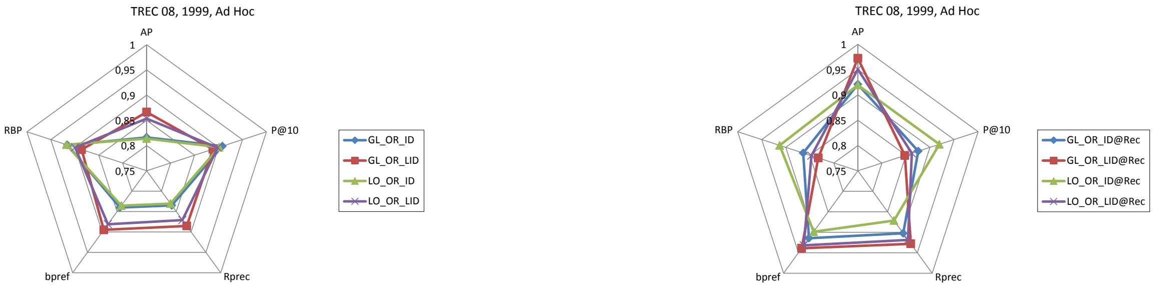
3.2. Analysis of the Properties of the Markov Measures

From the mathematical point of view we concentrate on the analysis of the invariant distribution of Markov chains. We will study how the shape of the invariant distribution depends on the relative position of the relevant documents. For instance, if in a ranked result list the order of some documents is modified we would like to approximate this change with a mathematical estimate which can predict the tendency of both the invariant distribution and the value of the measure. This estimate would be useful to tackle the significant problem represented by the cost of the experimental campaigns; if it were possible to judge only a small part of the great number of documents and to predict how the lack in the relevance judgements would affect the measure, then the employment of the resources would be less expensive.

From the experimental point of view we compare MP to the other evaluation measures (Average Precision (AP), P@10, Rprec, Rank-Biased Precision (RBP), and Binary Preference (bpref)); we conducted a correlation analysis and we studied its robustness to pool downsampling on the following data sets: TREC 7 Ad Hoc, TREC 8 Ad Hoc, TREC 10 Web, and TREC 14 Robust. As far as calibration of time is concerned, we used click logs made available by Yandex (Serdyukov et al. 2012). The full source code of the software used to conduct the experiments is available for download[1] in order to ease comparison and verification of the results.

Firstly, we computed the Kendall τ correlation (Kendall 1945) between MP and the performance measures of direct comparison. As a general trend MP tends not to have high correlations with the other evaluation measures (Figure 1a), indicating that it takes a different angle from them (users move forward and backward in the result list); but if we provide MP with the same amount of information AP has, i.e. we rescale MP by recall, the correlation with AP increases in almost all cases (Figure 1b). Then we analyse the effect of reducing the pool size on the absolute average performances and on the Kendall τ correlation. Concerning the absolute average performances MP shows a consistent behaviour over all

[1] `http://matters.dei.unipd.it/`



***(a)** Kendall τ correlation between different instantiations of MP and the other comparison measures.*

***(b)** Kendall τ correlation between different instantiations of rescaled MP and the other comparison measures.*

the collections: its absolute average values decrease as the pool reduction rate increases. If we consider the effect on the correlation,MP models tend to perform comparably to AP and, when provided with the same information about the recall base, they consistently improve their performances.

Finally, on the basis of the click logs, we can state that 21% of the observed transitions are backward, a fact that validates our assumption that a user moves forward and backward along the ranked list. Moreover, we compared the values of continuous-time MP and discrete-time MP, concluding that the continuous-time version depends heavily on the calibration of the holding times.

3.3. Development of a Web Prototype

As mentioned in Section 1, experimental evaluation generates huge amount of scientific data that need to be effectively and efficiently analysed, explored and understood. We plan to develop a prototype web application that allows the user not only to visualize the experimental results but to interact with them too. We will use VA tools (Keim et al. 2010), which integrate the user in the data mining process, and, through an efficient visualization of the information, he/she can interact with, modify and enhance the analysis of the data in order to detect the weak points of his/her system and to improve it.

Concerning the architecture of the prototype the Markovian user models and measures, as well as the data mining steps, will be performed using Matlab on the server side while, on the client side, an advanced web application will exploit information visualisation and VA techniques in order to engage and provide the user with an intuitive representation of the experimental results.

4. FUTURE WORK

Future works concern the investigation of alternative user models able to account also for the number of relevant/not relevant documents visited so far, the possibility of learning the transition probabilities of the Markov chain directly from click-logs, the calibration of time into MP, the investigation of the robustness of MP e.g. discriminative power (Sakai 2006), and the developmet of the web prototype.

5. ACKNOWLEDGMENTS

This work has been partially supported by ELIAS, a research networking programme (09-RNP-085) project (`http://elias-network.eu`) of the European Science Foundation (ESF).

REFERENCES

M. Angelini, N. Ferro, G. Santucci, G. Silvello, *VIRTUE: A Visual Tool for Information Retrieval Performance Evaluation and Failure Analysis*. In Journal of Visual Languages and Computing, volume 25, issue 4, pages 394–413, August 2014.

M. Ferrante, N. Ferro, M. Maistro, *Injecting User Models and Time into Precision via Markov Chains*. Proc. 37th Annual International ACM SIGIR, 2014, pages 597-606. ACM Press, New York, USA.

D. Keim, J. Kohlhammer, G. Ellis and F. Mansmann, *Mastering the Information Age Solving Problems with Visual Analytics*. Eurographics Association, Germany, 2010.

M. G. Kendall, *The Treatment of Ties in Ranking Problems*. Biometrika, 33(3):239–251, 1945.

A. Moffat and J. Zobel. Rank-biased Precision for Measurement of Retrieval Effectiveness. *ACM TOIS*, 27(1):2:1–2:27, 2008.

J. R. Norris. *Markov chains*. Cambridge University Press, UK, 1998.

C. J. Van Rijsbergen, *Information Retrieval*. Butterworth-Heinemann Newton, USA, 1979.

T. Sakai, *Evaluating Evaluation Metrics based on the Bootstrap*. SIGIR 2006, pages 525–532.

J. Savoy, *Statistical Inference in Retrieval Effectiveness Evaluation*. Information Processing and Management, Vol. 33, No. 4, pages 495–512, 1997.

P. Serdyukov, N. Craswell, and G. Dupret. WSCD2012: Workshop on Web Search Click Data 2012. In WSDM, pages 771–772. ACM, 2012.

Heterogeneous Information Access Through Result Composition

Horațiu Bota
School of Computing Science
University of Glasgow
18 Lilybank Gardens
Glasgow, UK
www.horatiubota.com
h.bota.1@research.gla.ac.uk

Modern search engines aggregate information from a variety of sources (e.g. images, videos) and return this information to users, merged into a single results page. Current aggregation techniques are limited to merging blocks of heterogeneous content into organic result rankings. We propose a new approach to search aggregation that takes into account result semantics and explicit searcher preferences in the form of *result composition*. Our findings suggest that result composition can be an effective search paradigm and can positively impact search behaviour in certain contexts.

Web search, search interfaces, user behaviour

1. INTRODUCTION

Over the past decades, we have seen an explosion of immediately accessible information. The first Internet search engine, Aliweb (Koster 1994), was released in 1994 and marked the beginning of a continuously expanding area of on-line activity: Web search. Today, with more than 80% of Internet surfers using search engines to satisfy their information needs (Shokouhi and Si 2011), search technology attracts notable attention from industry and research alike.

Modern search engines have started displaying results aggregated from heterogeneous information sources — also known as **verticals**, for example images, videos, news — in return to user queries. This search paradigm is primarily driven by the dynamics and volume of heterogeneous content on the Web, and is intended to support ambiguous information needs and varying user behaviour patterns (Zhou et al. 2012). Different approaches to aggregating information on the Web have been proposed and studied in prior work on *federated search*, *aggregated search* or *semantic web search*. In general, these approaches have been limited to searching multiple text collections simultaneously or merging results from different information sources within standard search results pages.

In light of these limitations, we argue that current approaches to aggregating results across different sources are insufficient and propose a new method of presenting users with a structured assembly of heterogeneous documents: *composite retrieval*. Instead of a ranked list of documents or a set of heterogeneous blocks, we want to provide users with semantically structured assemblies of documents, incorporated into a set of **composite information objects**.

Consider the following user information need “finding all information to plan a trip to Greece”. Answering this information need typically involves submitting several queries to gather information about airports and visa policies, to read on-line reviews about hotels, and to check the geographic proximity of places to visit. Current search engines aggregate results from multiple verticals, however, the presentation of search results is limited to heterogeneous blocks. As the web has made available a large variety of verticals, it is becoming important to return to users “organised” results, containing information extracted from different sources of information. Doing so will not only support users in complex search tasks, but also allow them to understand the diversity of the information space and select what matters to them most. Furthermore, composite retrieval on the Web can potentially promote exploratory user behaviour and move the

Learning and Development Ltd.
Proceedings of the 6th Symposium on
Future Directions in Information Access 2015

focus of search away from providing results to users' queries, towards providing answers to users' questions.

Our work focuses on returning to users results "organised" into what we refer to as a *composite information objects*, where individual objects contain results extracted from various sources (e.g. videos, images and blogs), each object focusing on different aspects of a user's query.

In broad terms, our research aims to answer the following research questions:

1. Can composition of results lead to better performance, in terms of traditional IR relevance, than ranked lists of results?
2. From the users' perspective, what results should composite objects contain? What functions do individual results play in a composite information object?
3. How does the presence of a composite information object on a traditional search results page influence user behaviour and perceived task workload? Can composite objects make search tasks easier for the user?

The following sections provide a brief overview of the existing literature on aggregation based Web search (Section 2), a summary of our current work (Section 3), and an outline of future directions for our efforts (Section 4).

2. BACKGROUND

The concept of responding to search queries by presenting a composition of items has been proposed and investigated in a number of recent papers (Guo et al. 2012; Tran et al. 2011; Zhao et al. 2011). Many of the above papers have provided contributions on the theoretical side, studying the complexity of evaluating queries with constraints, and proposing different algorithmic formulations. Other works have focused on building systems that perform composite retrieval under a number of different semantics and targeting specific application domains, such as on-line shopping or travel planning (Basu Roy et al. 2010; Xie et al. 2010).

In the context of heterogeneous information access, aggregated search is widely used by modern search engines. Aggregated search is the task of retrieving information from a variety of resources (or verticals) and merging it into a single interface (Arguello et al. 2009; Zhou et al. 2012). Aggregated search can be compared to federated search (Shokouhi and Si 2011) (also known as distributed information retrieval), which deals with merging result rankings from different search engines into one single ranking list. The main challenges in aggregated search and federated search are resource selection and result merging. The former deals with deciding which sources of information contain the most relevant results to a given query and the latter deals with selecting a subset of items from relevant sources and presenting them as results. In aggregated search, the most common result presentation strategy consists of merging blocks of heterogeneous results into ranked lists of organic results. Similar to aggregated search, selecting and organising results from heterogeneous sources is the main focus of composite retrieval. However, rather than presenting the results of each selected vertical as a block of homogeneous items, composite retrieval aims to present results into cohesive information objects, where each object contains heterogeneous items (retrieved from several verticals).

Understanding user search behaviour is a key component of modelling and evaluating search engine performance. In the context of aggregated search, user behaviour has been shown to differ significantly compared to the more traditional *ten blue links* environment. For example, in a study analysing click-through rates in an aggregated search scenario, Sushmita et al. (2010) found users click more on vertical results that are relevant to the task, shown higher in ranking and more visually salient. Diaz et al. (2013) mined users' mouse movement interactions from a commercial search engine log and found that different results presentation strategies create different biases with respect to user attention and browsing sequence. All previously mentioned studies investigate aggregated search scenarios in which heterogeneous content is displayed in blocks of items, embedded into the organic Web results list.

Unlike aggregated search, result composition merges results from different sources into singular information objects to be presented on the results page. From a presentation perspective, composite information objects are similar to entity cards (Navalpakkam et al. 2013; Lagun et al. 2014), which are shown on existing search results pages in response to ambiguous or entity-specific user queries. Entity cards are related to our work because they are instances of composite information objects: they contain heterogeneous results, extracted from different sources, assembled using various semantic retrieval techniques, and shown in response to ambiguous user queries. There is limited understanding of user behaviour in entity-card search: to our knowledge, only two prior studies examined the effects of entity cards on user interaction with search interfaces. With an eye and mouse tracking study, Navalpakkam et al. (2013) found that the flow of user attention on non-linear page layouts (with knowledge cards shown at the top-right corner of the SERP) is different from the widely believed, top-down linear examination of search results. In the context of mobile search, Lagun et al. (2014) performed a similar study in understanding how user attention is distributed between knowledge cards and web results in a mobile context.

3. CURRENT WORK

3.1. Retrieval Performance

To address our first research question, we adapted an existing composite retrieval framework (Amer-Yahia et al. 2013) to Web search. Due to the heterogeneous nature of the multi-vertical environment we explored, novel ways to model and estimate the various components of our proposed framework were developed. We used named entities (textual spots mapped to Wikipedia articles) to bridge the semantic gap between documents across verticals and developed algorithms to construct composite information objects. We applied our algorithms on a federated search test collection (Nguyen et al. 2012), which contained results from 108 search engines categorised into 11 different verticals.

Our results indicate that composite retrieval can significantly improve the performance over various current search paradigms, such as traditional “general web only” ranking, federated search ranking and aggregated search. The composite retrieval search paradigm we propose aims to promote a diverse information space for users to explore. For an exploratory task, rather than requiring searchers to issue multiple queries related to different aspects of their information need, issued to several vertical search engines, composite retrieval provides a unified page that consists of *relevant* objects focused on different aspects of a searcher’s query. Our results have implications for work in heterogeneous information access and diversity in IR. An in-depth analysis of results is presented in Bota et al. (2014).

3.2. User Perspectives

To address our second research question, we designed and ran a user study to analyse user-generated composite objects. Our main objective was to determine how composite objects are manually generated by searchers. In particular, our interest was to analyse composite objects with respect to their topical focus, content and user-assessed characteristics (i.e. relevance, cohesion, diversity).

Our results show that, firstly, there is an agreement between users on the topical focus of composite objects — namely that different users construct composite objects which focus on similar aspects of a given topic. Secondly, we observe that composite objects contain documents that play different roles. For instance, central documents (or pivots), are assessed by users as being more relevant than other document within the composite object, and reflect the object title, whereas ornament documents are less relevant but provide value to searchers through composition with pivots. Finally, our results suggest that no clear hierarchy of user-assessed object characteristics can be determined and that, although explicit relevance is crucial in search, composition of diverse results can generate additional value to users. An in-depth analysis of the results is available in Bota et al. (2015).

3.3. Effects on User Behaviour

In addition to aggregating results from different sources (e.g. images, video, news), modern search engines have started displaying complex information objects, or entity cards (ECs), on the results page. As mentioned in section 2, entity search cards are instances of composite information objects. Entity cards are intended to enhance search experience in several ways: ***(i)*** they help searchers navigate diversified results, ***(ii)*** provide a summary of relevant content directly on the results page and ***(iii)*** support exploratory search by highlighting relevant entities associated with a given user query. Because we wanted to understand the effect of result composition on user search behaviour, we designed and ran a large-scale crowd-sourced user study, with more than one thousand unique searchers, in which we studied the effect of entity cards and their properties (relevance, cohesion and diversity) on search behaviour and perceived task workload.

Our results suggest that the presence of ECs has a strong effect on both the way users interact with search results and their perceived task workload. The results of our investigation indicate that ECs have significantly different effects in simple versus complex tasks. Furthermore, by manipulating EC properties (*content*, *coherence* and *diversity*), we uncover different effects and interactions between card properties on measures of search behaviour and workload.

4. FUTURE WORK

In terms of future work, many open questions remain. With regard to our first research question, more rigorous evaluation metrics, tailored to non-linear search environments need to be developed in order to reliably investigate result composition performance.

Secondly, with regard to our second and third research questions, our work so far provides an extensive analysis of user behaviour in both a result composition scenario, and in an entity-card search scenario. However, because there is limited understanding of presentation strategies for composite information objects, we aim to investigate presentation optimisation strategies for result composition in Web search.

Given that mobile devices have become ubiquitous, and that Web search is increasingly prevalent on mobile devices, we intend to investigate the role result composition can play in mobile Web search. In particular, our objective is to understand whether user context (e.g. time, location, device size) can predict and explain the usefulness of composite information objects, as reflected by user engagement metrics, in mobile Web search.

5. CONCLUSION

As the Web has made available an enormous variety of textual and multimedia resources, people have started performing increasingly more complex search tasks, aimed at finding rich answers that require information extracted from various sources. To satisfy these complex information needs, modern search systems need to build solutions that aggregate information, taking into account users' intents and preferences. We argue that composition of results can provide users with a more structured approach to Web search. Our work so far suggests that returning composite information objects to users' search queries can not only provide a better search experience, in terms of traditional IR metrics, but can also positively impact user search behaviour and perceived task workload in certain contexts.

Many avenues for future work remain open. Fully understanding result composition requires the development of comprehensive evaluation metrics that take into account both the content and the presentation of composite objects. In addition, given the increasingly growing usage of mobile Web search, understanding the role composite objects play in various mobile contexts is crucial. We aim to address both these aspects in our future work.

Acknowledgements

The author would like to thank Dr. Ke Zhou and Prof. Joemon Jose for their invaluable support, and ELIAS (Evaluating Information Access Systems) for the scholarship to participate at ESSIR 2015 and the FDIA Symposium.

REFERENCES

Amer-Yahia, S., F. Bonchi, C. Castillo, E. Feuerstein, I. Méndez-Díaz, and P. Zabala (2013). Complexity and algorithms for composite retrieval. In *Proceedings of the 22nd international conference on World Wide Web companion*, pp. 79–80. International World Wide Web Conferences Steering Committee.

Arguello, J., F. Diaz, J. Callan, and J.-F. Crespo (2009). Sources of evidence for vertical selection. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, pp. 315–322.

Basu Roy, S., S. Amer-Yahia, A. Chawla, G. Das, and C. Yu (2010). Constructing and exploring composite items. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, SIGMOD '10, New York, NY, USA, pp. 843–854. ACM.

Bota, H., K. Zhou, and J. Jose (2015). Exploring composite retrieval from the users' perspective. In A. Hanbury, G. Kazai, A. Rauber, and N. Fuhr (Eds.), *Advances in Information Retrieval*, Volume 9022 of *Lecture Notes in Computer Science*, pp. 13–24. Springer International Publishing.

Bota, H., K. Zhou, J. M. Jose, and M. Lalmas (2014). Composite retrieval of heterogeneous web search. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, New York, NY, USA, pp. 119–130. ACM.

Diaz, F., R. White, G. Buscher, and D. Liebling (2013). Robust models of mouse movement on dynamic web search results pages. ACM CIKM '13, pp. 1451–1460.

Guo, X., C. Xiao, and Y. Ishikawa (2012). Transactions on large-scale data- and knowledge-centered systems vi. Chapter Combination Skyline Queries, pp. 1–30. Berlin, Heidelberg: Springer-Verlag.

Koster, M. (1994). Aliweb - archie-like indexing in the web. *Computer Networks and ISDN Systems 27*(2), 175 – 182. Selected Papers of the First World-Wide Web Conference.

Lagun, D., C.-H. Hsieh, D. Webster, and V. Navalpakkam (2014). Towards better measurement of attention and satisfaction in mobile search. In *ACM SIGIR '14*, pp. 113–122.

Navalpakkam, V., L. Jentzsch, R. Sayres, S. Ravi, A. Ahmed, and A. Smola (2013). Measurement and modeling of eye-mouse behavior in the presence of nonlinear page layouts. In *ACM WWW '13*, pp. 953–964.

Nguyen, D., T. Demeester, D. Trieschnigg, and D. Hiemstra (2012). Federated search in the wild: the combined power of over a hundred search engines. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pp. 1874–1878. ACM.

Shokouhi, M. and L. Si (2011). Federated search. *Foundations and Trends in Information Retrieval 5*(1), 1–102.

Sushmita, S., H. Joho, M. Lalmas, and R. Villa (2010). Factors affecting click-through behavior in aggregated search interfaces. In *ACM CIKM '10*, pp. 519–528. ACM.

Tran, Q. T., C.-Y. Chan, and G. Wang (2011). Evaluation of set-based queries with aggregation constraints. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, New York, NY, USA, pp. 1495–1504. ACM.

Xie, M., L. V. Lakshmanan, and P. T. Wood (2010). Breaking out of the box of recommendations: from items to packages. In *Proceedings of the fourth ACM conference on Recommender systems*, RecSys '10, New York, NY, USA, pp. 151–158. ACM.

Zhao, B., X. Lin, B. Ding, and J. Han (2011). Texplorer: Keyword-based object search and exploration in multidimensional text databases. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, New York, NY, USA, pp. 1709–1718. ACM.

Zhou, K., R. Cummins, M. Lalmas, and J. M. Jose (2012). Evaluating aggregated search pages. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pp. 115–124. ACM.

Temporal Information Retrieval Revisited

A Focused Study on the Web

Yue Zhao
Web Information Systems
Delft University of Technology
The Netherlands
y.zhao-1@tudelft.nl

Claudia Hauff
Web Information Systems
Delft University of Technology
The Netherlands
C.Hauff@tudelft.nl

Temporal information retrieval has been an active area of research for a number of years. Most existing works focus on the utility of temporal information in specific types of corpora (such as news archives), specific types of retrieval approaches, and, specific applications that may benefit from temporal information (such as timeline summarization). Underrepresented in existing works are studies that investigate the impact of temporal information analyses on the Web and Web documents. In this paper we (i) describe the research gaps we identified around Web-based temporal information analysis, and, (ii) present an overview of our first results and observations when studying *sub-document timestamping* on the Web.

Temporal information analysis, timestamping, sub-doucments, information diffusion

1. INTRODUCTION

One recurring theme in temporal information retrieval and analysis is the use of document creation timestamps for various tasks and applications, such as event detection [Döhling and Leser (2014)], document clustering [Alonso et al. (2009)] and the adaptation of retrieval algorithms to temporal queries [Li and Croft (2003)]. Determining the creation time of a Web document is challenging for a number of reasons: (1) document meta-data is generally unreliable, (2) public sources such as the Internet Archive (`https://archive.org/`) can only archive a small subset of the Web, and, most important of all, (3) the average Web document may change significantly over the course of its lifetime, in which case a single document creation timestamp is effectively only capturing when the Web document's location (URL) was first established on the Web, instead of when the document's content was created.

Existing works have largely addressed these challenges by relying on the document content itself to estimate a single creation date [de Jong et al. (2005); Kanhabua and Nørvåg (2009); Kumar et al. (2011); Chambers (2012); Ge et al. (2013)]. This approach has been shown to work well for document corpora that are static in nature, such as news corpora — each document is a single news article with few to no changes in content over time. Driven by the lack of research in *sub-document* timestamping (i.e. the labelling of document sentences and paragraphs with individual creation timestamps), we set out to investigate to what extent this simplifying assumption of a single creation timestamp also holds for Web documents.

This initial experiment (described in more detail in Section 2) led us to identify a number of perceived research gaps in Web-based temporal information analysis that we aim to investigate over the course of the next three years:

Web sub-document timestamping: Existing document timestamping algorithms [de Jong et al. (2005); Kanhabua and Nørvåg (2009); Kumar et al. (2011); Chambers (2012)] do not take the special nature of the Web into account, including its link structure and dynamic nature. Are we able to increase the accuracy of sub-document timestamping when utilizing this knowledge?

From News to Web corpora: Existing temporal retrieval approaches have been shown to outperform non-temporally aware approaches on news corpora [Berberich et al. (2010)]. To what extent does temporality aid in the retrieval of general Web documents?

Novel Applications: Assuming that sub-document timestamping is possible at Web scale, what types of novel applications or tasks can we solve? What kind of novel insights can we gain (e.g. about information diffusion [Yang and Leskovec (2010)])?

*© The Author(s). Published by BCS.
Learning and Development Ltd.
Proceedings of the 6th Symposium on
Future Directions in Information Access 2015*

In the remainder of this paper, we will first summarize our preliminary research on the timestamping of Web sub-documents (Section 2), followed by a broad outline of the research plan (Section 3) and an overview of open questions to be discussed at FDIA (Section 4).

2. SUB-DOCUMENT TIMESTAMPING

Our initial investigation (published in [Zhao and Hauff (2015)]) revolved around the assumption made in existing works utilizing document creation timestamps [Swan and Jensen (2000); Li and Croft (2003); Alonso et al. (2009); Jatowt et al. (2013); Döhling and Leser (2014)]: each document (no matter the corpus) is created at one point in time.

Although it is obvious that for Web documents this assumption generally does not hold, it is not yet known, to what extent the assumption is wrong. We designed an empirical analysis to investigate this issue, answering the following research questions:

RQ1 To what extent do Web documents consist of sub-documents created at different times?

RQ2 What is the timespan between the oldest and most recent sub-document of a document?

RQ3 What fraction of the current document has been created in each version (a version corresponding to a particular timestamp)?

RQ4 To what extent can the timestamp of each sub-document be predicted?

2.1. Data Set

Due to the preliminary nature of our study, we investigated a small sample of Web documents in depth, specifically the 11,075 relevant documents of the ClueWeb12 corpus (`http://www.lemurproject.org/clueweb12.php/`, TREC Web topics 201-300). This choice ensured that each investigated Web document is at least relevant to some information need.

2.2. Processing Pipeline

We first divided each of these ClueWeb12 documents into sub-documents (each sub-document is simply a paragraph). In order to learn when each sub-document was created, we crawled all historic versions of the document stored before 2012 (the crawl date of ClueWeb12) that are available at the Internet Archive — overall, 64% of our documents were captured by the Internet Archive with on average 17 historical versions.

Based on these historical versions, we determined the earliest version in which each sub-document occurred and assigned the Internet Archive crawl date of that version as sub-document creation timestamp. First, we extract all textual contents and divide them into sub-documents by HTML tags (e.g. `<p>` and `<div>`). Subsequently, these sub-documents are compared to sub-documents in the historical versions. If two sub-documents have a high similarity, they are treated as being the same and sub-documents in ClueWeb12 documents are timestamped by their earliest occurrence in the Internet Archive.

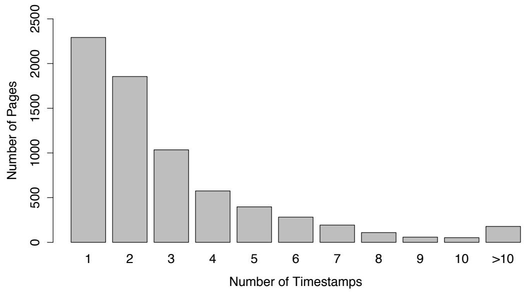


Figure 1: *Overview of the number of documents containing content created at different points in time.*

Lastly, we also extracted more than 20 features (such as the length and position of sub-documents and the number and value of temporal expressions in sub-documents) from each sub-document in order to train & test our timestamp classifier. Based on the ground truth sub-document timestamps, we created five timestamp classes by dividing them in balance, which means that each class has ∼55K instances. For example, the time intervals of the first two classes are $A = [0, 20.5], B = (20.5, 311.5]$, which means that the sub-documents in class A have been created no later than 20.5 days before the ClueWeb12 document's crawl date, while sub-documents in class B have been created between 20.5 and 311 days before the document's crawl date respectively. We aim to predict each sub-document's timestamp class correctly.

2.3. Results

The results shown in Figure 1 indicate that the majority of Web documents have indeed more than one creation timestamp (answering **RQ1**): 67% of Web documents consist of sub-documents with at least two different creation timestamps, with only a small percentage (less than 4%) having more than 8 creation timestamps.

When considering the difference in days between a document's oldest and most recent sub-document creation timestamp (**RQ2**) we find a surprisingly large gap in Figure 2: the median difference is 400

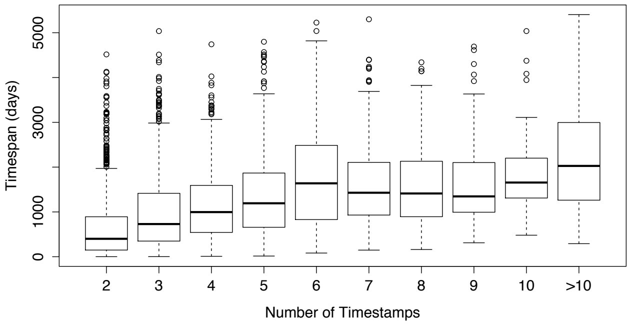


Figure 2: *The document set is partitioned according to the number of creation timestamps (documents with a single creation timestamp are ignored). Shown is the difference (in days) between the oldest and most recent creation timestamp.*

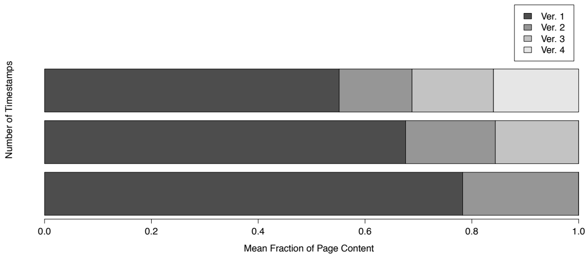


Figure 3: *Overview of content created at different points in time for documents with 2, 3 or 4 creation timestamps. Each bar shows the mean fraction of content available at each creation timestamp.* Ver. 1 *indicates the content created at the oldest timestamp,* Ver. 2 *the content created at the second oldest timestamp and so on.*

days, i.e. 50% of our investigated Web documents contain content created more than one year apart.

Lastly, we turned our attention to the amount of content created at each point in time (**RQ3**), restricting our analysis to those documents with 2, 3 or 4 creation timestamps as shown in Figure 3. Evidently, most of a Web document's content is created initially; the more creation timestamps a document has, the lower the percentage of initially created content. Moreover, we find that the average contents updated in each time are similar, which has also been found in earlier studies [Fetterly et al. (2004); Ntoulas et al. (2004)].

Having analysed the extent of a Web document's content being created at different points in time, we also experimented with the prediction of the correct timestamp class (**RQ4**). In a 5-class setup and our 20+ features we were able to classify 64% of all sub-documents correctly, significantly better than our baseline classifier which considered only the temporal expression within each sub-document for classification purposes (39% accuracy).

These results indicate that utilizing the creation times of sub-documents (instead of a single creation time per document) are likely to have a significant effect on Web retrieval tasks that utilize this type of temporal information.

3. RESEARCH PLAN

Based on these encouraging results, we developed a research plan for the coming three years, revolving around temporal information analysis on the Web.

We aim to investigate the following research themes:

Web sub-document timestamping: Our first goal is to drastically scale up the analysis of sub-document timestamping along the lines of the preliminary experiments. Ideally, instead of investigating 11,000 ClueWeb12 documents, we investigate *all 733 Million* ClueWeb12 documents. This change in magnitude will allow us to also effectively investigate novel features for the prediction of sub-document timestamps such as the link structure *between* sub-documents and time-series based features (due to the dynamic nature of the Web). Additionally, we aim to move beyond standard classification towards a more fine-grained approach using sequential labeling methods [Lafferty et al. (2001); Dietterich (2002)] which can exploit the relationships among a document's sub-documents.

From News to Web Corpora: While past works have mostly investigated news corpora (partially due to their unambiguity in creation timestamps), we aim to extend these works by employing proposed (as well as newly developed) retrieval approaches to Web corpora, investigating the impact of sub-document timestamps on retrieval effectiveness.

Novel Applications: The envisioned large-scale nature and at the same time fine-grained analysis of sub-document timestamping will allow us to consider novel applications, such as investigating the effect of the rise (or decline) of particular portals on information diffusion on the Web. Information diffusion on the Web might be influenced by specific websites. For example, programmers prefer to talk about programing problems on StackOverflow nowadays rather than writing their problems on their blogs as before.

4. OPEN QUESTIONS

There are a number of open questions, that would be particularly interesting to discuss during FDIA.

(1) The main limitation of our work is the restricted accuracy of the sub-document timestamps we generate with our Internet Archive-based

methodology. While for popular Web documents the crawling frequency is high, unpopular documents are crawled infrequently and thus, the timespan between the crawling time and the real creation time of a sub-document may be large. Are there ways to improve the methodology to generate more accurate timestamps for unpopular Web documents?

(2) We expect that in contrast to other research areas where humans achieve a very high accuracy (e.g. face recognition or image content labelling), sub-document timestamping to be very challenging for human labellers. This leads to the questions of (i) how to measure when a prediction is accurate enough, and, (ii) how to determine whether or not there is a hidden ceiling for the prediction accuracy of sub-document timestamping?

(3) In previous works on temporal information retrieval, the temporal information is leveraged as filters or additional conditions. Is it possible to combine temporal and other features (e.g. in a learning to rank setting) as innate features rather than additional restrictions?

5. ACKNOWLEDGEMENT

We thank the ESF ELIAS for the awarded scholarship for participating in ESSIR 2015 and the FDIA Symposium.

REFERENCES

Alonso, O., M. Gertz, and R. Baeza-Yates (2009). Clustering and exploring search results using timeline constructions. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pp. 97–106. ACM.

Berberich, K., S. Bedathur, O. Alonso, and G. Weikum (2010). *A language modeling approach for temporal information needs*. Springer.

Chambers, N. (2012). Labeling documents with timestamps: Learning from their time expressions. In *ACL '12*, pp. 98–106.

de Jong, F., H. Rode, and D. Hiemstra (2005). Temporal language models for the disclosure of historical text. Royal Netherlands Academy of Arts and Sciences.

Dietterich, T. G. (2002). Machine learning for sequential data: A review. In *Structural, syntactic, and statistical pattern recognition*, pp. 15–30. Springer.

Döhling, L. and U. Leser (2014). Extracting and aggregating temporal events from text. In *Proceedings of the companion publication of the 23rd international conference on World wide web companion*, pp. 839–844. International World Wide Web Conferences Steering Committee.

Fetterly, D., M. Manasse, M. Najork, and J. L. Wiener (2004). A large-scale study of the evolution of web pages. *Software – Practice & Experience 34*(2), 213–237.

Ge, T., B. Chang, S. Li, and Z. Sui (2013). Event-based time label propagation for automatic dating of news articles. In *EMNLP '13*, pp. 1–11.

Jatowt, A., C.-M. Au Yeung, and K. Tanaka (2013). Estimating document focus time. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pp. 2273–2278. ACM.

Kanhabua, N. and K. Nørvåg (2009). Using temporal language models for document dating. In *Machine Learning and Knowledge Discovery in Databases*, pp. 738–741.

Kumar, A., M. Lease, and J. Baldridge (2011). Supervised language modeling for temporal resolution of texts. In *CIKM '11*, pp. 2069–2072.

Lafferty, J., A. McCallum, and F. C. Pereira (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Li, X. and W. B. Croft (2003). Time-based language models. In *Proceedings of the twelfth international conference on Information and knowledge management*, pp. 469–475. ACM.

Ntoulas, A., J. Cho, and C. Olston (2004). What's new on the Web?: the evolution of the Web from a search engine perspective. In *Proceedings of the 13th International Conference on World Wide Web*, New York, NY, USA, pp. 1–12. ACM.

Swan, R. and D. Jensen (2000). Timemines: Constructing timelines with statistical models of word usage. In *KDD Workshop on Text Mining*, pp. 73–80.

Yang, J. and J. Leskovec (2010). Modeling information diffusion in implicit networks. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pp. 599–608. IEEE.

Zhao, Y. and C. Hauff (2015). Sub-document timestamping of web documents. In *Proceedings of the 38th international ACM SIGIR conference on Research & development in information retrieval*. ACM.

Syntactic and Semantic Structures for Relation Extraction

Duc-Thuan Vo
Laboratory for Systems, Software and Semantics (LS^3)
Ryerson University, Toronto, ON, Canada
ducthuan.vo@ryerson.ca

Ebrahim Bagheri
Laboratory for Systems, Software and Semantics (LS^3)
Ryerson University, Toronto, ON, Canada
bagheri@ryerson.ca

This study proposes to employ syntactic and semantic knowledge from the rich relations within a tree kernel structure for relation extraction. The underlying idea is that different tree kernels with a variety of representations of the available linguistic information will improve the performance of detecting useful pieces of information expressed in a sentence. Applying clause-based rules, clustering algorithms, and bootstrapping on them will help increase the performance of relation extraction. As outlined in this paper, we plan to conduct experiments on recent Information Extraction corpuses and compare the results with the state of the art.

Syntactic. Semantic. Tree kernel. Clause-based relation. Clustering algorithms, Bootstrapping

1. INTRODUCTION

Relation extraction (RE) is one of the challenging tasks in information retrieval. The goal of relation extraction is to discover the relevant segments of information in large numbers of textual documents such that they can be used for structuring data. RE aims at discovering various semantic relations in natural language text. It has been applied in many information retrieval tasks such as question answering. For instance answering the question “Who is the President of the United States?” would require a structure where the entity “Barrack Obama” would have the relation “the President of” with another entity “United States”.

Some of the existing research in RE obtains a shallow semantic representation of natural lanaguage text in the form of verbs or verbal phrases and their arguments (Bankko et al., 2008; Fader et al., 2011; Wu et al., 2010). Other approaches such as WOEparse (Wu et al., 2010), OLLIE (Mausam et al., 2012), and ClausIE (Corro et al., 2013) use dependency parsing for relation extraction. Each of these approaches makes use of various heuristics to obtain propositions from dependency parsers. Furthermore, bootstrapping (Xu et al., 2007; Xu et al., 2010; Etzioni et al., 2005; Bunescu et al., 2007) has been applied in relation extraction, which does not need a large amount of predefined labels on the training data. It starts from a small set of n-ary relation instances as “seeds”, in order to automatically learn pattern rules from parsed data, which would then be used to extract new instances of relations. Such ER systems learn extraction patterns from dependency trees automatically and systematically induce rules with different complexities. Moreover, several research works have exploited unsupervised methods for relation extraction. They have tried to address this challenge by building on the latent relation hypothesis which states that pairs of words that co-occur in similar contexts tend to have similar relations (Turney, 2008; Rosenfeld et al., 2007; Akbik et al, 2012; Akbik et al., 2014). The authors exploited features using dependency tree to discover relations by clustering entity pairs. Cluster vector space model (pattern) is applied by using the k-mean algorithm and cosine similarity is used to measure distances.

However, existing research face some limitations such as:

1. (P1) Using dependency trees may result in incoherent and uninformative extractions in cases where the extracted relation phrase has no meaningful interpretation. For example, given a sentence “*They recalled that Nungesser began his career as a precinct leader.*", the words *recalled* and *began* are linked together that will create an incoherent relation based on dependency tree-based methods. This will limit maximum recall or may lead to a significant drop of precision at higher points of recall as reported in (Mausam et al., 2012; Wu et al., 2010; Felder et al., 2012; Corro et al., 2013).
2. (P2) Several earlier works such as (Mausam et al., 2012; Wu et al., 2010, Felder et al., 2011) try to apply heuristic rules with Hidden Markov Models (HMMs) or Conditional Random Fields (CRFs)-based sequence labeling for RE. CRF-based approaches are state of the art and they

Proceedings of the 6th Symposium on Future Directions in Information Access 2015

have yielded high performance in sequence learning tasks. However, the supervised nature of CRF relies on a fairly large amount of training data which must be annotated by humans (Mausam et al., 2012; Corro et al, 2013).

3. (P3) In the work by Xu et al. (Xu et al., 2007; Xu et al., 2010), bootstrapping is applied with predefined rules to train relations based on a dependency tree. However, this approach results in low performance when used on unobserved new domains due to the high likelihood of extracting incorrect rules from the dependency tree during the bootstrapping process.

In order to propose a framework that can address the above three challenges, we have identified the Tree kernel representation to be a solid foundation for our work as it allows us to capture a variety of information including semantic concepts, words, POS tags, shallow and full syntax, dependency parsing, and discourse trees (Xu et al., 2013; Saleh et al., 2014; Zhou et al., 2010; Nguyen et al., 2009; Bunescu et al., 2005). In this study, we will deal with the above three challenges by exploiting the tree kernel structure as follows:

- We will use linguistic knowledge from grammar clauses of the English language to detect relations in rich syntactic and semantic structures for addressing P1. Heuristic rules are applied to obtain proposition relations from the rich tree structure. The rich tree structure includes POS tags, shallow and full syntax, dependency parsing, and discourse trees from the tree kernel that can automatically determine the relations in a sentence. We will use heuristic rules to obtain proposition relations from the rich tree structure.

- In order to address P2, we will model a rich semantic relation tree structure as a vector space model from different tree kernels based on the latent relation hypothesis (Turney, 2008). This representation can compute the similarity of arguments (entity pairs) of relation by comparing the distribution over observed patterns. We then apply clustering methods to find clusters of entity pairs that share similar patterns that can be assumed to represent a relation.

- Finally, we will extend bootstrapping methods by analyzing features from rich syntactic and semantic structures from discourse trees in order to address P3.

2. RELATED WORK

The task of relation extraction was first introduced in the Message Understanding Conference (MUC-6). Since then, a number of techniques have been proposed for this task such as feature vector-based methods and tree kernel-based methods (Xu et al., 2013; Zhou et al., 2010; Nguyen et al., 2009; Bunescu et al., 2005; Vo et al., 2012). Open Information Extraction (OIE) was first presented by Banko et al. (2007) by not being restricted to a pre-specified list of relations in RE. More recent work in Open IE (Akbik, 2009; Wu et al., 2010; Fader et al., 2011; Mausam et al., 2012) have received significant attention. Most of these research work use a shallow semantic representation or dependency parsing in the form of verbs or verbal phrases and their arguments (Banko et al., 2007; Wu et al., 2010; Fader et al., 2011). Mausam et al. (2012) present an improved system called OLLIE, which relaxes the previous systems' constraints that relation words are mediated by verbs, or relation words that appear between two entities. OLLIE creates a training set which includes millions of relations extracted by REVERB (Fader et al., 2011) with high confidence. OLLIE learns relation patterns from the dependency path and lexicon information. Relations that matched the extracted patterns are extracted.

In unsupervised and weakly supervised learning, several authors have built on the latent relation hypothesis which states that pairs of words that co-occur in similar patterns tend to have similar relations (Turney, 2008; Rosenfeld et al., 2007; Akbik et al, 2012; Akbik et al., 2014). These authors exploited features from the dependency tree for discovering relations by clustering entity pairs. Cluster vector space model (pattern) is often applied by using the k-mean algorithm and cosine similarity is used to measure distances. By applying bootstrapping (Xu et al., 2007; Xu et al., 2010; Etzioni et al., 2005; Bunescu et al., 2007), Xu et al., (2007) and Xu et al., (2010) have presented a framework for the extraction of relations. They do not need a large number of predefined labels on the training data. The bootstrapping-based model starts from a small set of n-ary relation instances as "seeds", in order to automatically learn pattern rules from the seed data, which can then extract new relation instances.

As mentioned earlier, the use of dependency trees (Mausam et al., 2012; Corro et al., 2013; Xu et al., 2013) might limit maximum recall or may lead to the drop of precision at higher points of recall due to incoherent and uninformative extractions. Also, RE methods that have employed bootstrapping (Xu et al, 2007; Xu et al., 2010) are limited in their application to new domains due to their focus on relations that are domain specific. We believe that the tree kernel can be a rich syntactic and semantic structure that includes semantic concepts, words, POS tags, shallow and full syntax, dependency parsing and discourse tree (Xu et al., 2013; Saleh et al., 2014; Zhou et al., 2010; Nguyen et al., 2009), which can help to improve the performance when identifying pieces of relation information in a sentence. We suggest that the tree kernel has potential for improving the performance of ER techniques. Our work aims to augment the tree kernel structure with additional semantic, e.g. named entities concepts and syntactic, e.g. explicit relation nodes (Moschitti, 2006; Zhou et al., 2010; Saleh et al., 2014) for relation extraction as outlined in the following section.

3. OVERVIEW OF THE PROPOSAL APPROACH

The common definition of the RE task is a function from a sentence to a set of triples, such as < *E*1*, R, E*2 >, where *E*1 and *E*2 are entities (noun phrases) and *R* is a relation between the two entities. Several RE systems extract specific relations for prespecified named entity types (Zhou et al., 2010; Nguyen et al., 2009; Bunescu et al., 2005). For instance, *R.MarriedTo(E1.Per, E2.Per)* or *R.LocatedAt(E1.Org, E2.Loc)*. Open Information Extraction (Open IE) (Banko et al., 2007; Corro et al., 2013; Wu and Weld, 2010; Fader et al., 2011; Mausam et al., 2012), a type of RE, aims to extract general relations for two entities. The idea of Open IE is to extract a diverse range of relations and avoid the need for a specific training relation set. For example, *(Tom, married, Marry)* or *(Tom, studies, Computer Science)*. In our work, we propose the following contributions:

3.1. Contribution 1: Tree kernels and clause-based relations

A relation candidate can consist of words before, between, or after the relation pair, or the combination of two consecutive positions. With tree kernel, both learning and classification rely on the inner-product between instances. Tree kernels avoid extracting explicit features from parse trees by calculating the inner product of the two trees, and instead they rely on the common substructure of two trees. We will exploit clauses of the English language to detect relations in rich semantic tree structure. A clause is a part of a sentence that expresses some coherent piece of information; it consists of one subject (S), one verb (V), and optionally an indirect object (O), a direct object (O), a complement (C), and one or more adverbials (A).

Given a sentence "*Obama, the president of the United States, was born in Hawaii on August 4, 1961*", Figure 1 (a) shows the shortest dependency tree path (SDTP) between "*Obama*" and the "*United States*". Additionally, Figure 1 (b) shows a tree kernel with an *R* node added based on the unlexicalized Grammatical Relation Centered Tree (Croce et al. (2011). And, *R* node is as a relation in tree structure. In this example, if a clause structure such as subject-verb-object is considered and *R* is bound to a verb, then relations like *S:Obama; V:the president; O:the United States*, *S:Obama; V:was born; O:Hawaii and S:Obama*; *V:was born; O:on August 4 ,1961* can be extracted.

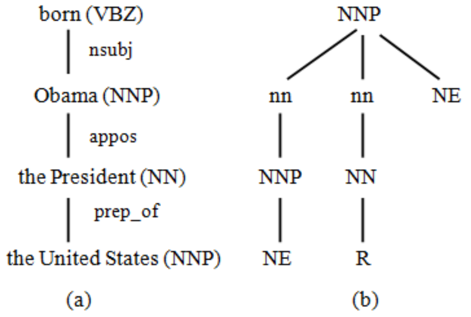


Figure 1: *(a) The shortest dependency tree path (SDTP); (b) Tree structure with "R" added.*

To model the RE problem according to the above example, we will first construct a rich tree structure for a sentence based on the tree kernel. We then gather clauses which exist in the sentence. For each clause, we will determine the set of coherent derived-clauses based on the dependency path, e.g., (*Obama, was born, in Hawaii*) and (*Obama, was born, on August 1961*) from (*Obama, was born*). Finally, we will use heuristics rules to determine, and supervised learning methods such as SVM to classify the proposition relations.

3.2. Contribution 2: Tree kernels and clustering algorithms

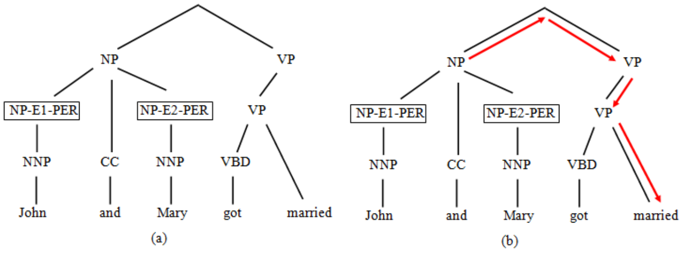


Figure 2: *(a) The shortest dependency tree path (SPT); (b) Predicate-linked: SPT and the rich parse tree structure.*

Current techniques (Turney, 2008; Akbik et al, 2012; Akbik et al., 2014) exploit features from the dependency tree for discovering relations by clustering entity pairs. We choose not to use the dependency path for word extraction due to challenges mentioned above. We will construct a rich semantic-relation tree structure as a vector space model based on different tree-based kernels. We will also discover relations in each sentence by clustering entity pairs. For example, both sentences "*John and Mary got married.*" and "*John and his wife Mary joined Microsoft*." show the relation *MarriedTo* between entity pairs "*John*" and "*Mary*". We characterize each relation based on a set of common patterns. As an example, Predicate-linked (Figure 2.b) of the sentence "*John and Mary got married*" and tree structure with "R" added (Figure

3.b) of the sentence "*John and his wife Mary joined Microsoft.*" have the similar patterns. The vector space model based on the extracted patterns will follow the latent relation hypothesis (Turney, 2008). We then apply clustering methods to find clusters of entity pairs that share similar patterns representing a specific relation.

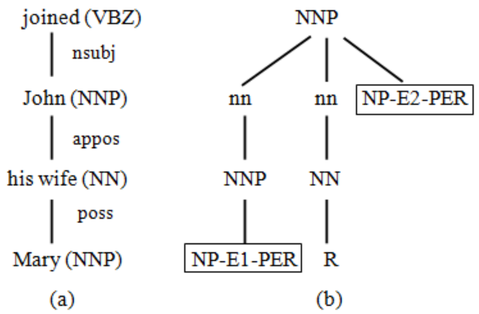


Figure 3: *(a) The shortest dependency tree path (SDTP); (b) Tree structure with "R" added.*

3.3. Contribution 3: Tree kernels and bootstrapping

Xu et al. presented bootstrapping (Xu et al., 2007; Xu et al., 2010) with pre-defined rules to train relations from a dependency tree. Their approach shows low performance in new unobserved domains due to its reliance on a specific corpus. Therefore, in our third contribution, we will use a tree kernel to address the limitations in the dependency tree method through combining it with self-training methods. Our approach will start with some extracted patterns containing potential relations and a small set of relation instances as "seed" in order to train new patterns. The extracted patterns will be based on existing clauses (clauses mentioned in Section 3.1.) in the sentence that will not be limited to a small set of relation types.

For instance, let us consider two relation types *MarriedTo(E1.Per, E2.Per)* and *LocatedAt(E1.Org, E2.Loc)*. The relation *MarriedTo* would need to be associated with two entity pairs *(E1.Person, E2.Person)* and a set of common relation words such as <*"married", "lover", "..."*>. Furthermore, the relation *LocatedAt* is associated with entity pairs *(E1.Org, E2.Loc)* and a set of common relation words like <*"located", "is at", "..."*>. The self-training methods rely on the RlogF metric whereby those patterns that have more words related with relation instance seeds will receive a higher score (Thelen et al., 2002; Patwardhand et al., 2007). In Figure 4, the extracted patterns P1 and P3 receive high scores for the *MarriedTo* relation and will hence be added to seed of relation *MarriedTo*. The seed of this relation type will be updated with new common words such as "*married*" and "*wife*". Also, the extracted patterns P2 and P6 are added in seed of relation *LocatedAt* with new common word such as "*located at*". Therefore, the system will self improve in the next iteration. By using the self-training method, we will build a new relation list by continuously adding new information to the seeds of the relations.

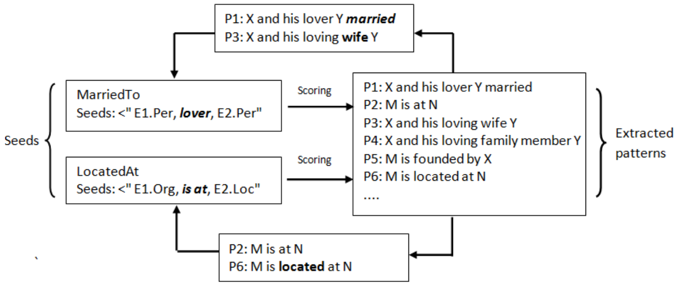

P1: X and his lover Y ***married***
P3: X and his loving **wife** Y
MarriedTo
Seeds: <" E1.Per, ***lover***, E2.Per"
LocatedAt
Seeds: <" E1.Org, ***is at***, E2.Loc"
Seeds
Scoring
Scoring
P1: X and his lover Y married
P2: M is at N
P3: X and his loving wife Y
P4: X and his loving family member Y
P5: M is founded by X
P6: M is located at N
....
Extracted patterns
P2: M is at N
P6: M is **located** at N

Figure 4: *Self-training process in two relation types.*

4. EVALUATION PLAN

There are three main datasets that are widely used for the evaluation of RE techniques, namely REVERB[1], OLLILE[2] , and ACE[3]. REVERB provides 1,000 tagged training sentences and 500 test sentences. REVERB also provides extracted relations and instance confidence values for the 500 test sentences. OLLIE has a test set which has 300 sentences as well as 900 extracted triples. Finally, ACE RDC 2004 corpus contains 451 documents and 5,702 positive relation instances. It redefines seven entity types, seven major relation types and 23 relation subtypes. Most of the state of the art RE systems perform experiments on these corpuses.

In terms of the state of the art performance, Fader et al., (2011) focus on efficiency by restricting syntactic analysis to part-of-speech tagging and chunking and obtained about precision of 57% and recall of 64% on the REVERB dataset. Mausam et al., (2012) use dependency parsing and various heuristics to obtain propositions relation. They archived around precision of 63% with 600 extracted relations from OLLIE dataset. Corro et al., (2013) also made use of dependency parsing combined with a set of sentence clauses the use various heuristics to obtain propositions from the dependency parses. They archived a precision of 59% with 3,000 extracted relations in REVERB. Xu et al., (2013) proposed multiple SVM models with dependency tree kernels for relation extraction on REVERB and OLLIE datasets, and achieved F-measures of 78.1% in REVERB and 79.3% on OLLIE. Zhou et al., (2010) explored diverse features through a linear kernel and with Support Vector Machines (SVM), and achieved an F-measure of 77.8% in ACE RDC 2004 corpus. Jiang et al., (2007) evaluated the effectiveness of

[1] http://reverb.cs.washington.edu
[2] http://knowitall.github.io/ollie
[3] http://www.ldc.upenn.edu

different feature subspaces with different complexities and obtained the best F-measure of 71.5% on the seven relation types of the ACE RDC 2004 corpus.

We will use these RE corpuses for experiments in our three contributions and compare with state of the art approaches. REVERB and OLLIE will be employed in our first contribution due to not being restricted to a prespecified list of relations. In order to compare with the state of the art such as Xu et al. (2010) and Akbik et al. (2014) in contributions 2 and 3, the ACE RDC 2004 will be used for experiments. We will also use the Stanford parser for analyzing syntactic and semantic structures to be combined with tree kernel (Moschitti et al., 2006).

5. CONCLUDING REMARKS

In this paper, we introduce our proposal for addressing three challenges in RE. We believe that by adding rich syntactic and semantic relation structures to tree kernels, we will be able to improve the state of the art in relation extraction. Our core contribution is to enrich kernel trees with crucial syntactic and semantic information combined with techniques such as clause-based rules, clustering algorithms, and bootstrapping for relation extraction.

ACKNOWLEDGEMENT

We thank ELIAS (Evaluating Information Access System) ESF Research Networking Programme for the awarded scholarship for participating in ESSIR 2015 and FDIA Symposium.

REFERENCES

Akbik, A., Michael, T., Boden, C. (2014) Exploratory Relation Extraction in Large Text Corpora. In Proceedings of *COLING 2014*.

Akbik, A., Visengeriyeva, L.(2012). Unsupervised Discovery of Relation and Discriminative Extraction Patters. In Proceedings of *COLING 2012*.

Banko, M., Cafarella, M.J., Soderland, S., Broadhead, M., Etzioni, O. (2007). Open Information Extraction from the Web. In Proceedings of *IJCAI 2008.*

Bunescu, R., Mooney, R.J. (2005). Subsequence kernels for relation extraction. In Proceedings of *NIPS 2005*.

Bunescu, R., Mooney, R. (2007). Learning to extract relations from the web using minimal supervision. In Proceedings of *ACL 2007*.

Corro, L.D., Gemulla, R. (2013). ClausIE: Clause-Based Open Information Extraction. In Proceedings of *WWW 2013*.

Etzioni, O., Cafarella, M., Downey, Doug., Popescu, A.M., Shaked, T., Soderland, S., Weld, D., Yates, Alexander. (2005). Unsupervised named-entity extraction from the web: An experimental study. *Artificial Intelligence*, vol. 165, pp. 91 – 134.

Fader, A., Soderland, S., Etzioni, O. (2011) Identifying Relations for Open Information Extraction. In Proceedings of *EMNLP 2011.*

Jiang, J., Zhai, CX. (2007). A systematic exploration of the feature space for relation extraction. In Proceedings of *NAACL-HLT 2007*.

Mausam, Schmitz, M., Bart, R., Soderland, S. (2012). Open Language Learning for Information Extraction. In Proceedings of *EMNLP 2012.*

Moschitti A. (2006) Making tree kernels practical for natural language learning. In Proceedings of *EACL 2006*.

Nguyen, T.V.T, Moschitti, A. (2009). Convolution kernels on constituent, dependency and sequential structures for relation extraction. In Proceedings of the *EMNLP 2009*.

Patwardhan S., Riloff E. (2007). Effective Information Extraction with Semantic Affinity Patterns and Relevant Regions. In Proceedings of *EMNLP-CoNLL 2007*.

Rosenfeld, B., Feldman, R. (2007). Clustering for unsupervised relation identification. In Proceedings of the *CIKM 2007*.

Saleh, I., Moschitti, A., Nakov, P., Marquez, L., Joty, S. (2014). Semantic Kernels for Semantic Parsing. In Proceedings of *EMNLP 2014*.

Thelen M., Riloff E. (2002). A Bootstrapping Method for Learning Semantic Lexicons using Extraction Pattern Contexts. In Proceedings *EMNLP 2002*.

Turney, P. (2008). The latent relation mapping engine: Algorithm and experiments. *Journal of Artificial Intelligence Research*, 33(1):615–655.

Vo, D. T., Ock, C. Y. (2012). Extraction of Semantic Relation Based on Feature Vector from Wikipedia. In *PRICAI 2012: Trends in Artificial Intelligence*, Springer Berlin Heidelberg.

Wu, F., Weld, D.S. (2010). Open Information Extraction using Wikipedia. In Proceedings of *ACL 2010*.

Xu, F., Uszkoreit, H., Li, H. (2007). A Seed driven Bottom up Machine Learning Framework for Extracting Relations of Various Complexity. In Proceedings of *ACL 2007*.

Xu, F., Uszkoreit, H., Krause, S., Li, Hong. (2010). Boosting Relation Extraction with Limited Closed-World Knowledge. In Proceedings of *COLING 2010*.

Xu, Y., Kim, M.Y., Quinn, K., Goebel, R., Barbosa, D. (2013). Open Information Extraction with Tree Kernels. In Proceedings of *NAACL-HLT 2013*.

Zhou, G., Qian, L., Fan, J. (2010). Tree kernel-based semantic relation extraction with rich syntactic and semantic information. *Information Sciences*, vol. 180, 2010.

The Opinionated Recommender

Stephen Bradshaw
College of Engineering and Informatics
National University of Ireland Galway, Ireland

Insight Centre for Data Analytics
National University of Ireland Galway, Ireland
s.bradshaw1@nuigalway.ie

Recommender Systems (RSs) are devices that are used to filter data to combat information overload and provide time saving measures to the user. While RSs have traditionally been done using a content or collaborative based approach, recent times have seen a surge in alternative approaches to try and alleviate some of the traditional problems found there such as the filter bubble, matrix scarcity and cold start issues. Many of these new approaches attempt to lever new sources to provide more accurate recommendations and offset some of these issues. In this paper we will outline some of the current flaws and propose a hypothetical system that will exploit external sources to improve upon the state of the art.

Recommender Systems, Topic Extraction, Social Media

1. INTRODUCTION

Traditionally recommendations have been approached in two ways, either on a content basis or through collaborative means. Content based recommendations is done by comparing the attributes of an item and recommending items of a similar nature. This system has been used very successfully for years by companies like Amazon [4]. Issues with content based recommender systems (CBRs) is that of *overspecialisation* [13]. This is when a recommender becomes so attuned to the customer profile that is only recommends things that the customer already has knowledge of; and fails to produce content that is novel or contains serendipitous value. [1] notes that serendipity has two facets, the degree of surprise for the user and usefulness. In addition RSs can suffer from *matrix scarcity*, where there are very few ratings on items/people from which to base recommendations on.

Collaborative recommender systems (CFs) recommend items based on their similarity to those known to be of interest to similar users. They do this by constructing a summary of interests or profile and find other users with similar profiles in order to make further recommendations. The result of using a community of people is that it can offset the issue of *overspecialisation* and can introduce elements of serendipity and novelty in the recommendations, however traditional issues with this approach include *cold start* problems [10]. The cold start problem is where new users have not provided enough feedback from which to base new recommendations on. User feedback is itself a big problem in the area of RSs as it can be difficult to get users to interact with a system that does not produce instantaneous and accurate results. In addition, CF approaches are prone to create scenarios where the user is exposed to a *filter bubble* [7]; where a user only hears back things that support her own views, and are not exposed to conflicting points of view.

Our aim is to investigate and design a recommender system that does not fall prey to the issues stated above. We believe that through the application of text content analysis, and harvesting social media we can improve upon the quality of returned information and improved information assimilation can be made throug improved presentation on areas of interest. Improvements can come in the form of highlighting sections that have garnered a lot of attention or perhaps through a graph format. To foster a degree of serendipity we aim to incorporate views found on social media platforms (like Reddit and boards.ie). Our work focuses on identifying topics in a text and linking those to topics expressed in opinions. In Section 2 we will talk about some recently proposed solutions to the problems stated and their shortcomings. Section 3 will deal with some approaches to deal with returning personalised content. In Section 4 we will outline a hypothetical

Proceedings of the 6th Symposium on Future Directions in Information Access 2015
Proceedings of FDIA

approach that we believe is capable of tackling these issues as well as dealing with traditional flaws with recommender systems.

2. TRUST BASED RECOMMENDER SYSTEMS (TBRS)

One approach to offset some of the issues stated above is proposed by [6]; who argue that CFs treat all profiles as independent entities and fail to acknowledge that they might contain an element of interconnectivity amongst the users. They coin their model *social trust ensemble* which aims to get at the core of CFs principles: which is that we will accept the recommendations of a friend over that of a stranger. To prove their hypothesis they create a use case which recommends films to users. They use Epinions.com as source from which to make their predictions. Epinoins is a site where people can give ratings of 1 - 5 on items. In addition the users provide a trusted list of associates, which the authors use to identify trusted connections. They can then increase the coverage of the recommendations by incorporating the trusted friend list into the process. This they state increases the coverage of the system and offsets the sparsity issue. Finally, they use a probabilistic factor analysis model (PFAM) to produce the final results. The strength of this approach is that PFAM is not computationally expensive, and so can be extended to a larger dataset. Their system shows that coverage can be increased using fewer recommendations, however it requires the creation of a trusted list. One could argue that they are merely moving the sparsity problem to that of the trusted friends list. Without having a sufficiently large set of trusted users the system will succumb to the same issue.

Another system that uses opinions as a source to create a trust based recommender system is that one proposed in [5]. They argue that their system is built on a *web of trust* where greater weight is given to recommendations that are coming from trusted associates. One strength of their approach is that it reduces the changes of malicious recommendations by favouring those that come from a trusted source. Like [6] they feel that a trust based recommendation system will alleviate the sparsity problem and new user issue.Iit is worth noting that Epinions.com is significantly smaller database 83,509 ratings than some other movie review database. Eachmovie contains 2, 811, 983 ratings and Movielens has 1,000,209 ratings [6]. We argue that while there are merits for incorporating trust into recommendations, existing databases that contain both trust metrics and reviews are more sparsely populated than standard movie review databases and thus are not a viable solution to the issue.

[5] conduct a study on the strengths of incorporating trust into recommender systems through a study of Cyworld. Cyworld is a Korean friend site of a similar nature to Facebook. The authors conducted a user study that has 42 members select their favourite skine. A skine is a profile picture accessory that can be purchased to alter the appearance of your profile page. They determined 'trust' levels by measuring the amount of interaction between two users. Interaction is determined from number of messages left on profile walls. The results from their study show that by incorporating additional social data into standard CF approach one can improve the quality of the recommendations.

3. PERSONALISED SEARCH

An approach that aimed to achieve a similar outcome to our own was conducted by Teevan [11]. Her approach automatically created two profiles for the user, one with previous searches and the other on pages visited. She experimented with trying various additional inputs to improve results, namely; processing emails exchanges, calender dates and documents stored on the users computer. She evaluated her work on a test group and found that user satisfaction can be achieved through personalising the search results returned.

An additionally approach that aimed to gauge how well general opinion can be factored into a personalised recommender was done through the aid of Amazons Mechanical Turk. The Mechanical Turk is a market place run by Amazon that matches workers to employers in the performance of simple repetitive tasks. General tasks can include subjective analysis like which colour is nicer. Workers who perform these tasks are referred to as turk workers. Experimentation was performed on doing two tasks *taste matching* and *taste grokking*. The first applied CF techniques on groups of turk workers and formed groups of people with similar interests, from which to determine new items of interest for a user. The second was to give the turk workers a number of sample items and see if they can recommend additional items that the user would like. The strengths of the authors approach is that they created a novel personalisation system that can make subjective decisions dynamically. They also demonstrated that harvesting the opinion of the crowd can be very beneficial in designing a recommender. Another strength was that the system required little user knowledge. In the next section we will sketch our own proposed approach which also aims to incorporate opinion as a factor to improve recommendations.

4. THE OPINIONATED RECOMMENDER

To utilise content found on social sites we propose applying information retrieval methods to data found on chat forums (Boards.ie / Reddit). Our assumption is that people express views on contemporary issues like those expressed by journalists there. In addition we assume that these views are influenced by content found in daily newspapers. We propose a system that mines news websites for the topics contained within. In our proposed system a user inputs to the system a URL to a news article on sport that she finds interesting. Topic are mined from this article in order to identify topics of interest. Topics are a useful item to determine from stories as they contain a large amount of semantic information.

Topic extraction can potentially be performed using natural language processing techniques. An example for such is Part Of Speech Tagging(POS). POS is where nouns, adjectives and verbs are identified and used for syntax analysis [2]. A sliding window approach [12] can be used to see which noun noun, noun adjective pairings are occurring most frequently and they are in turn used to ascertain the intended topic of the text . Rousseau et al [9] propose an interesting variation on the standard tf-idf approach [8] called *tw-idf* where instead of counting term occurrences *tf*, they count term co-occurrences *tw* and graph them as edges on a node, which might offer an interesting platform to expand upon for determining topics and their interconnectivity. Open questions in this regard include; can the number of topics in an article be determined from density of clusters, and how dense does a cluster have to be before it is considered a topic ? Finally is there a means of identifying one term that can accurately express the topic of the assembled words?

An external source such as DBpedia is then used to obtain additional information on the topics. The same process is applied to the chat forums to see if the topics identified are being discussed there. Once the topics have been identified we propose graphing the topics discussed in the chat forums and applying distance metrics to see what additional topics are most closely related to the users' preference topics. A vector is then created storing the additional topics weighted by their proximity to known topics of interest. The vector might be used to inform a decision tree on the users interest; where topics may be the leaves of the tree and the presence of a certain combination of leaves will indicate a user's interest or otherwise. Newspaper conglomerations (such as Google News) are then crawled to see if new hitherto unknown articles can be found. These articles are returned to the user in a hierarchical order of interest. In addition, hot topics or topics that are producing a lot of 'chat' can be highlighted so the user gains an extra level of depth on the issues.

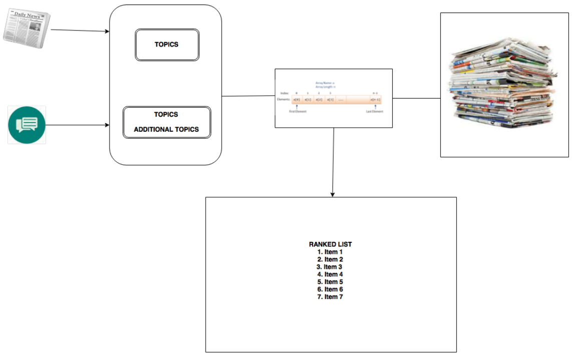


Figure 1: *System Architecture*

5. DISCUSSION

Traditional approaches to recommender systems have reached the zenith of their ability and there is been a marked increase in the number of researchers looking for additional ways to improve them [9] [11] [12]. External sources such as social media sites are seen as having big potential to be utilised to improve upon recommendations. In addition making recommendations informed by user feedback has shown to be insufficient on its own [3] as the context of a user's situation influences what numeric feedback she will apply. The user may rate a mediocre film more generously if she had watched a very bad film prior and inversely a good film may only receive a middling rating if the film the rater had watched prior was superb. We argue that content based measures can be exploited in the form of content analysis, and that a collaborative-like approach can increase the diversity of recommendation. We propose a system that is based on opinion expressed in chat forums, which can be exploited by applying information retrieval approaches to evaluate the information found there and use it to augment RSs. Our approach aims to avoid some of the standard RS issues such as *cold start, information sparsity* and add *serendipity* to the recommendations. Future work includes building a system that incorporate the above outlined approaches and evaluating whether such an approach can improve the coverage as well as satisfaction of a user.

ACKNOWLEDGEMENTS

I would like to thank my supervisor Matthias Nickles for his assistance and feedback. I also would like to thank the reviewers for their insights and helpful pointers.

REFERENCES

[1] Mouzhi Ge, Carla Delgado-Battenfeld, and Dietmar Jannach. Beyond accuracy: evaluating recommender systems by coverage and serendipity. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 257–260. ACM, 2010.

[2] Thomas L Griffiths, Mark Steyvers, David M Blei, and Joshua B Tenenbaum. Integrating topics and syntax. In *Advances in neural information processing systems*, pages 537–544, 2004.

[3] Jonathan L Herlocker, Joseph A Konstan, Loren G Terveen, and John T Riedl. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)*, 22(1):5–53, 2004.

[4] Greg Linden, Brent Smith, and Jeremy York. Amazon. com recommendations: Item-to-item collaborative filtering. *Internet Computing, IEEE*, 7(1):76–80, 2003.

[5] Fengkun Liu and Hong Joo Lee. Use of social network information to enhance collaborative filtering performance. *Expert systems with applications*, 37(7):4772–4778, 2010.

[6] Hao Ma, Irwin King, and Michael R Lyu. Learning to recommend with explicit and implicit social relations. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):29, 2011.

[7] Tien T Nguyen, Pik-Mai Hui, F Maxwell Harper, Loren Terveen, and Joseph A Konstan. Exploring the filter bubble: the effect of using recommender systems on content diversity. In *Proceedings of the 23rd international conference on World wide web*, pages 677–686. ACM, 2014.

[8] Juan Ramos. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, 2003.

[9] François Rousseau and Michalis Vazirgiannis. Graph-of-word and tw-idf: new approach to ad hoc ir. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 59–68. ACM, 2013.

[10] Guy Shani and Asela Gunawardana. Evaluating recommendation systems. In *Recommender systems handbook*, pages 257–297. Springer, 2011.

[11] Jaime Teevan, Susan T Dumais, and Eric Horvitz. Personalizing search via automated analysis of interests and activities. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 449–456. ACM, 2005.

[12] Ville H Tuulos and Henry Tirri. Combining topic models and social networks for chat data mining. In *Proceedings of the 2004 IEEE/WIC/ACM international Conference on Web intelligence*, pages 206–213. IEEE Computer Society, 2004.

[13] Sung-Shun Weng and Mei-Ju Liu. Feature-based recommendations for one-to-one marketing. *Expert Systems with Applications*, 26(4):493–508, 2004.

Two-dimensional point set pattern matching with horizontal scaling

Antti Laaksonen
Department of Computer Science
University of Helsinki
ahslaaks@cs.helsinki.fi

This paper focuses on two-dimensional point set pattern matching with horizontal scaling. Given a dataset S of n points and a pattern P of m points, the task is to find occurrences of P in S. The pattern may be horizontally scaled using a constant value. This problem is relevant in symbolic music information retrieval when each point is interpreted as a musical note and the pattern is a melody that is searched for. The best known general algorithm for the problem works in $O(n^2m)$ time. In this paper we show that the 3SUM problem can be reduced to this problem, and present a new algorithm that works in $O(n^2)$ time on typical inputs.

1. INTRODUCTION

A natural problem in symbolic music information retrieval is to search for occurrences of a melody in a musical score. This problem has a geometric interpretation where each musical note is a point in the plane. Using this interpretation, the x coordinate of a point corresponds to the onset time of a note and the y coordinate corresponds to the pitch. In addition, horizontal scaling of the pattern should be allowed to adapt to tempo differences between the melody in the query and in the occurrences.

Let us now define the problem more formally. The *dataset* is an array of n points $S[1], S[2], \ldots, S[n]$, and the *pattern* is an array of m points $P[1], P[2], \ldots, P[m]$. Each point is a pair (x, y) where x and y are real number coordinates. We use a dot syntax for accessing the values. For example, $P[3].x$ is the x coordinate of the third point.

Given two points $p_1 = (x_1, y_1)$ and $p_2 = (x_2, y_2)$, we define their order lexicographically, i.e., $p_1 \leq p_2$ if either $x_1 < x_2$ or $x_1 = x_2$ and $y_1 \leq y_2$. Throughout the paper, we assume that S and P are already sorted in the lexicographic order. Furthermore, we use the notations $p_1 + p_2 = (x_1 + x_2, y_1 + y_2)$ and $p_1 - p_2 = (x_1 - x_2, y_1 - y_2)$.

Our goal is to find the indices in the point set from which a *occurrence* of the pattern begins. An occurrence consists of m indices $1 \leq i_1 < i_2 < \ldots < i_m \leq n$ and fulfils the following conditions. First, there is a *scaling factor* $\alpha > 0$ such that $S[i_{k+1}].x - S[i_k].x = \alpha(P[k+1].x - P[k].x)$ for each $k = 1, \ldots, m-1$. Second, we require that $S[i_{k+1}].y - S[i_k].y = P[k+1].y - P[k].y$ for each $k = 1, \ldots, m-1$.

There are two major difficulties in the problem. First, there may be gaps between occurrence points in the dataset. For this reason, standard string matching algorithms cannot be used for solving the problem. Second, partial pattern occurrences with different scaling factors cannot be combined, which makes it difficult to design any dynamic programming style algorithm for the problem.

The problem was first discussed by Romming and Selfridge-Field (2007) in the context of partial pattern matching using geometric hashing. Lemström (2010) defined the problem similar to this paper and gave an $O(n^2m \log n)$ time algorithm. A more efficient algorithm with time complexity $O(n^2m)$ was proposed by Laaksonen (2013). A related problem to the problem is one-dimensional point set pattern matching with scaling, where the difference is that the y coordinate of each point is constant. This problem can also be solved in O(n^2m) time, as shown by Rezende and Lee (1995).

The structure of the rest of the paper is as follows: In Section 2, we show how the 3SUM problem can be reduced to this problem. In Section 3, we present a new algorithm that solves the problem in $O(n^2)$ time under certain assumptions. Finally, in Section 4, we present our conclusions.

*© The Authors. Published by BCS Learning and Development Ltd.
Proceedings of the 6th Symposium on Future Directions in Information Access 2015*

2. 3SUM REDUCTION

The 3SUM problem can be defined as follows: Given an array A of integers, is it possible to choose three integers $a, b, c \in A$ such that $a+b+c=0$? Gajentaan and Overmars (1995) noticed that many problems in computational geometry are at least as difficult as 3SUM. Next we show that two-dimensional point set pattern matching with horizontal scaling also belongs to this group of problems.

The 3SUM problem can be easily solved in $O(n^2)$ time, and it was conjectured for a long time that no $o(n^2)$ time solution exists. However, a recent paper by Grønlund and Pettie (2014) refutes the conjecture by presenting an algorithm that solves 3SUM slightly faster, in $O(n^2(\log\log n)^{5/3}/(\log n)^{2/3})$ time.

We can solve 3SUM in $O(n \log n + f(n))$ time assuming that we can use a subroutine X that solves two-dimensional point set pattern matching with horizontal scaling in $f(n)$ time. First we sort the numbers in A in $O(n \log n)$ time. After this, we check if there is a solution where $a = b$, $a = c$ or $b = c$. For every number $k \in A$, we check if A also contains $-2k$. This can be done in $O(n \log n)$ time using binary search. Another special case is the trivial solution $a = b = c = 0$ when A contains 0.

From this point on, if a solution a, b, c exists, the numbers are distinct and nonzero. There are two subproblems: either two numbers are positive and one is negative, or two numbers are negative and one is positive. We solve the subproblems separately using our subroutine X. For the first subproblem we construct set S_1, and for the second subproblem we construct set S_2. Pattern P always consists of three points: $(1,\,1)$, $(2,\,2)$ and $(3,\,3)$.

Then for each number $k \in A$ we add one or two points to the datasets. If $k > 0$, we add points $(k,\,1)$ and $(k,\,3)$ to S_1 and point $(k/2,\,2)$ to S_2. If $k < 0$, we add points $(-k,\,1)$ and $(-k,\,3)$ to S_2 and point $(-k/2,\,2)$ to S_1. Now, P can be found in S_1 or S_2 exactly when there are three distinct numbers $a, b, c \in A$ such that $a + b + c = 0$.

For example, suppose that $A = [3, 5, -7, 4, -2]$. Now S_1 consists of points $(1, 2)$, $(3, 1)$, $(3, 3)$, $(3.5, 2)$, $(4, 1)$, $(4, 3)$, $(5, 1)$ and $(5, 3)$. Figure 1 shows the points in S_1 and an occurrence of P that consists of points $(3, 1)$, $(3.5, 2)$ and $(4, 3)$ with scaling factor 0.5. This occurrence corresponds to the solution $a = 3, b = -7, c = 4$ for the 3SUM problem.

Next we prove the construction in the general case where a, b and c are distinct nonzero numbers.

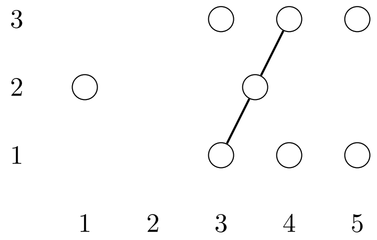


Figure 1: *An instance of the construction used in the 3SUM reduction. Each circle is a dataset point, and a pattern occurrence is shown.*

First, assume that $a + b + c = 0$, a and b are positive, $a < b$, and c is negative. Now set S_1 contains points $(a, 1)$, $(b, 3)$ and $(-c/2, 2) = ((a + b)/2, 2)$, so pattern P appears in S_1 with scaling factor $(b - a)/2$. Correspondingly, if a and b are negative and c is positive, set S_2 contains points $(-a, 1)$, $(-b, 3)$ and $(c/2, 2) = ((-a - b)/2, 2)$.

Then, assume that S_1 contains an occurrence of P. The occurrence consists of points $(z + \alpha, 1)$, $(z + 2\alpha, 2)$ and $(z + 3\alpha, 3)$. Thus, A contains values $z+\alpha$, $z+3\alpha$ and $-2(z+2\alpha)$, and the sum of the values is 0. Correspondingly, if S_2 contains an occurrence of P, A contains values $-(z + \alpha)$, $-(z + 3\alpha)$ and $2(z + 2\alpha)$, again with sum 0.

3. ALGORITHM

In this section we present an algorithm that solves two-dimensional point set pattern matching with horizontal scaling in $O(n^2)$ time under two assumptions. First, we assume that all coordinates of the points are integers. Second, we assume that there is a contant c such that the horizontal distance between any two consecutive points in the dataset is at most c.

The idea in the algorithm is to limit the number of possible scaling factors that need to be checked. For a fixed scaling factor, all pattern occurrences can be found in $O(nm)$ time using a technique presented by Ukkonen (2003). We show that under the above assumptions, there are only $O(n/m)$ possible scaling factors, and using this fact we can construct an algorithm that solves the problem in $O(n^2)$ time.

Listing 1 shows the structure of the algorithm. Note that for simplicity in the presentation, we assume that each pattern point has a distinct x coordinate. However, the algorithm could be modified to handle general patterns by grouping pattern points that have equal x coordinates.

Algorithm 1

```
 1: $e \leftarrow \infty$
 2: for $k \leftarrow 2, \ldots, m$ do
 3:     $e \leftarrow \min(e, P[k].x - P[k-1].x)$
 4: end for
 5: for $z \leftarrow 1, \ldots, S[n].x - S[1].x$ do
 6:     $\alpha \leftarrow z/e$
 7:     if $\alpha(P[m].x - P[1].x) \leq S[n].x - S[1].x$ then
 8:         $P' \leftarrow P$
 9:         for $k \leftarrow 2, \ldots, m$ do
10:             $P'[k].x \leftarrow P'[k-1].x + \alpha(P[k].x - P[k-1].x)$
11:             $Q[k] \leftarrow 1$
12:         end for
13:         for $i \leftarrow 1, \ldots, n$ do
14:             $c \leftarrow 1$
15:             for $k \leftarrow 2, \ldots, m$ do
16:                 while $Q[k] < n$ and
17:                         $S[Q[k]] - S[i] < P'[k] - P'[1]$ do
18:                     $Q[k] \leftarrow Q[k] + 1$
19:                 end while
20:                 if $S[Q[k]] - S[i] = P'[k] - P'[1]$ then $c \leftarrow c+1$
21:             end for
22:             if $c = m$ then print($i$)
23:         end for
24:     end if
25: end for
```

3.1. Analysis

First, on lines 1–4, the algorithm calculates the minimum horizontal distance between two consecutive points in the dataset. Then, on lines 5–25, the algorithm goes through all possible scaling factors. After selecting the scaling factor, on lines 8–23, the algorithm uses a technique similar to Ukkonen (2003) for searching for the pattern occurrences. Next we will focus on the condition on line 7 that limits the number of scaling factors to be checked.

The algorithm is based on the following observation: in any pattern occurrence the scaled horizontal distance between some two consecutive pattern points is at most $cn/(m-1)$. The reason for this is that the horizontal distance between the first and last dataset point is at most cn. Therefore there has to be two points in the pattern whose scaled horizontal distance is at most $cn/(m-1)$, because otherwise the scaled horizontal distance between the first and the last pattern point would be more than cn.

To calculate the number of the possible scaling factors we use the fact that all coordinates are integers. Thus, the minimum scaled horizontal distance between two pattern points is an integer between 1 and $\lfloor cn/(m-1) \rfloor$. The time complexity of the algorithm is $O(n^2)$ because there are $O(n/m)$ possible scaling factors to check and each check works in $O(nm)$ time.

3.2. General case

It would be tempting to try to limit the number of possible scaling factors also in the general case. However, the following construction shows that if there can be arbitrary gaps between dataset points (thus a constant c is not involved), there can be $\Theta(n)$ pattern occurrences with distinct scaling factors. Given integers n and m, the construction produces an instance of the problem with $n' \geq n$ dataset points, exactly m pattern points, and $\Theta(n')$ pattern occurrences with distinct scaling factors.

The database consists of q point groups, each having a total of $2m-1$ points. The groups are numbered $1, \ldots, q$. The value q is selected so that it is the minimum number for which $n' = q(2m-1) \geq n$. Each note in group x is a pair $(p(x)t(y), x)$ where $y = 1, \ldots, 2m-1$; $p(k)$ is the k'th prime number, $t(1) = 0$ and $t(k) = 2^{k-2}$ if $k > 1$. The pattern consists of pairs $(t(x), 1)$ where $x = 1, \ldots, m$.

For example, suppose that $n = 20$ and $m = 4$. In this case $2m-1 = 7$, $q = 3$ and $n' = 21$. The dataset consists of three groups of points:

1. $(0,1)$, $(2,1)$, $(4,1)$, $(8,1)$, $(16,1)$, $(32,1)$, $(64,1)$
2. $(0,2)$, $(3,2)$, $(6,2)$, $(12,2)$, $(24,2)$, $(48,2)$, $(96,2)$
3. $(0,3)$, $(5,3)$, $(10,3)$, $(20,3)$, $(40,3)$, $(80,3)$, $(160,3)$

The pattern is $(0,1), (1,1), (2,1), (4,1)$ and there are 4 pattern occurrences in each group. In group 1 the scaling factors are 2, 4, 8 and 16, in group 2 they are 3, 6, 12 and 24, and in group 3 they are 5, 10, 20 and 40.

In the general case, in each group there are m pattern occurrences. The total number of pattern occurrences with distinct scaling factors is qm, and $n' = q(2m-1) < 2qm$, thus the number of scaling factors is $\Theta(n')$.

4. CONCLUSIONS

In this paper we showed that the 3SUM problem can be reduced to two-dimensional point set pattern matching with horizontal scaling. This suggests that it is unlikely that the problem could be solved significantly more efficiently than in $O(n^2)$ time.

In addition, we presented a new algorithm for the problem that works in $O(n^2)$ time. The algorithm assumes that all coordinates in the input are integers and there is a constant that limits the distance between consecutive points in the dataset. These assumptions hold in many applications. For example, in music information retrieval, note onset times can be seen as integers and there cannot be long rests between consecutive notes.

It would be interesting to know if the problem could be solved in $o(n^2m)$ time in the general case. The one-dimensional point set pattern matching problem could be a good starting point because it captures the special case where all y coordinates have a constant value.

5. ACKNOWLEDGEMENTS

The participation of the author in the ESSIR 2015 summer school and in the FDIA 2015 symposium has been supported by the Helsinki Doctoral Programme in Computer Science and the ELIAS ESF Research Networking Programme.

REFERENCES

Gajentaan, A. and Overmars, M. "On a class of O(n^2) problems in computational geometry," *Computational Geometry*, 5(3), pp. 165–185, 1995.

Grønlund, A. and Pettie, S.: "Threesomes, degenerates, and Love Triangles," available at `http://arxiv.org/abs/1404.0799`, 2014

Laaksonen, A.: "Efficient and simple algorithms for time-scaled and time-warped music search," *CMMR 2013*

Lemström, K.: "Towards more robust geometric content-based music retrieval," *ISMIR 2010*

Rezende, P. and Lee, D.: "Point set pattern matching in d-dimensions," *Algorithmica*, 13(4) pp. 387–404, 1995

Romming, C. and Selfridge-Field, E.: "Algorithms for polyphonic music retrieval: the Hausdorff metric and geometric hashing," *ISMIR 2007*

Ukkonen, E., Lemström, K. and Mäkinen, V.: "Geometric algorithms for transposition invariant content-based music retrieval," *ISMIR 2003*

Topic-centric Classification of Twitter User's Political Orientation

Anjie Fang[1], Iadh Ounis[2], Philip Habel[2], Craig Macdonald[2] and Nut Limsopatham[2]
University of Glasgow, UK
[1]*a.fang.1@research.gla.ac.uk,* [2]*{firstname.secondname}@glasgow.ac.uk*

We aim to classify people's voting intentions by the content of their Tweets about the Scottish Independence Referendum (hereafter, IndyRef). By observing the IndyRef dataset, we find that people not only discussed the vote, but raised topics related to an independent Scotland including oil reserves, currency, nuclear weapons, and national debt. We show that the views communicated on these topics can inform us of the individuals' voting intentions ("Yes" vs. "No"). In particular, we argue that an accurate classifier can be designed by leveraging the differences in the features' usage across different topics related to voting intentions. We demonstrate improvements upon a Naive Bayesian classifier using the topics enrichment method. Our new classifier identifies the closest topic for each unseen tweet, based on those topics identified in the training data. Our experiments show that our proposed Topics-Based Naive Bayesian classifier improves accuracy by 7.8% over the classical Naive Bayesian baseline.

1. INTRODUCTION

Twitter emerged as an especially popular platform during the IndyRef held in 2014. We propose a technique to analyse the voting intentions of users, based on data mining and machine learning approaches. The general approach we propose could also be used to understand users' voting intentions in other major elections. To analyse voting intentions, we capture two months of Twitter data related to the IndyRef. To form a ground truth, we label users based upon hashtags appearing in their tweets, and we verify the reliability of this approach using the users' followee networks. After removing the hashtags from these tweets, we then focus on the remaining terms, treating each term as a feature. However, the referendum created an evolving discourse, with different topical themes (such as *oil*, *currency*, and *debt*), which make the accurate classification of users' voting intentions more challenging. For instance, the word "change" is indicative of a "No" voter in the *currency* topic, and of a "Yes" voter in the *nuclear weapons* topic. That is, there was a significant discussion over whether Scotland would need to "change" its currency if it obtained independence, while the "Yes" camp purported that the nuclear arsenal base could "change" in an independent Scotland. The dichotomy of the term "change" in indicating voting intentions across different topics highlights the main benefit of our approach. Indeed, this paper contributes the use of topical clusters to identify the topic of discussion in a tweet and subsequently it leverages this topic to classify the user's voting intention. Our approach, called *Topics-Based Naive Bayesian* (TBNB) demonstrates marked improvements over a classical Naive Bayes (NB) classification baseline.

2. BACKGROUND AND RELATED WORK

Cohen and Ruths (2004) demonstrated that classification of political orientation was still a difficult problem and that the earlier result in Al Zamal et al. (2004) was exaggerated since it used easily classifiable political data. We focus on the content of tweets to classify the users' voting intentions. We use as a starting point a classical Naive Bayesian (NB) classifier. Since the number of features can be very large, we use several feature selection approaches in Mladenic and Grobelnik (1999). Each selection approach ranks and selects F informative features based on the training data. Of course, not every selected feature will appear in the unseen test tweets.

3. TOPICS-BASED NAIVE BAYESIAN

The IndyRef discussions on Twitter revolved around a number of topics, for which people's opinions usually reflected their vote intentions. Let us continue the example of the word "change" usage in Section 1. The difference in usage of "change" across different topics is high. Furthermore, the conditional probability of "change" in the "Yes" category is higher than in the “No” category in the “currency” topic. Typically, the feature selection approaches just select features with higher differences between categories. If a feature differs between topics (e.g. “change”), it will be treated as different features in the TBNB model. Thus TBNB can capture term dependencies between topic and user voting intentions. Our TBNB classifier leverages both the features’ dissimilarities across topics and in the categories. In the training step, the topics are first detected by Latent Dirichlet Allocation (LDA). For each topic, a corresponding probability table is produced, where each feature has two associated conditional probabilities related to the two possible voting intentions (“Yes”/“No”). Consequently, during the training step, we produce as many feature tables as the number of used topics. In the testing step, we treat a user as a virtual document and this document contains the users’ tweets. For each tweet in the user’s virtual document, the topic that is closest to the tweet’s content is selected. Terms in an unseen tweet are then examined using the probability table generated during the training step for the topic with which this tweet is associated. In this way, terms in different tweets are treated differently based on their associated topics, and the TNBN classifier applies, for each unseen tweet, those features that were learned from the corresponding topic. Note that the feature selection approaches can naturally be applied to the TBNB classifier. For example, if F is set to 1000, the top 1000 features learned from each topic are selected.

4. REFERENDUM DATA AND EXPERIMENTS

Our IndyRef dataset was collected from Twitter by searching for a number of referendum-specific hashtags and keywords using the Twitter API from August 1, 2014 to September 30, 2014. In our dataset, certain “Yes” hashtags (e.g. #YesBecause) were associated with a “Yes” vote, and “No” hashtags (e.g. #NoBecause) with a “No” vote. To generate our ground truth, we assume that if a user’s tweets are only tagged by “No” hashtags, this user is labeled as a “No” voter. Similarly, if a user’s tweets contain only “Yes” hashtags, this user is labeled as a “Yes” supporter, favoring independence. Using this method, we find 5326 “Yes” users and 2011 “No” users. Together these 7337 users account for more than 420K tweets. After labelling, all “Yes” and “No” hashtags are removed from their original tweet text. The resulting tweets constitute our classification dataset. Without the hashtags, the classification task is naturally more challenging, but importantly, the resulting generalisable classifier does not require the presence of hashtags. We verify our ground-truth’s reliability using the users’ followee networks. In particular, If a user mainly follows Conservative politicians (“No” campaign supporters), this person is likely to be a “No” voter. If a user follows Scottish National Party politicians (“Yes” campaign supporters), their vote intention is more likely to be “Yes”. We then examined the networks of the 7337 users in our dataset, and identified who these users follow among the 536 public Twitter accounts corresponding to Members of the British or Scottish Parliaments. We find that, of the 7337 users, 87% can be verified into “Yes” or “No” voters, demonstrating that our ground-truth produced by the hashtags labeling method is reasonable and reliable.

We use our IndyRef dataset to compare the performances of the NB and TBNB classifiers. We vary the number of selected features F and the deployed feature selection approach for both NB and TBNB. We also vary the number of topics T in the TBNB classifier. We use a 10-fold cross validation process over the 7337 users and use accuracy to measure the performance. Our results show that all TBNB classifiers markedly outperform the NB baseline when F ranges from 10K to 50K. The highest accuracy of TBNB (90.4%) is achieved when applying the weighted odds ratio feature selection approach with T=10 and F =30K, while the accuracy of the baseline is 82.6%. In an additional experiment aiming to check the generalisation of our conclusions, we obtained similar results using a different IndyRef dataset (collected from different period) with the same aforementioned T and F values.

5. CONCLUSIONS AND FUTURE WORK

We classified the users’ voting intentions on Twitter during the IndyRef. We noted that the users tended to focus their discussions on topics, reflecting their voting intentions. We proposed to enrich the Naive Bayes classifier by leveraging the underlying topics covered in the tweets. Our proposed approach leverages the difference of the features across the topics and voting categories to increase the classification confidence. Our results demonstrate the effectiveness of our resulting TBNB classifier on two datasets. In the future, we plan to analyse the effect of the evolving discussions on the users’ voting intentions over time.

6. ACKNOWLEDGEMENTS

We thank ACM SIGIR for the awarded scholarship for participating in ESSIR 2015 and the FDIA Symposium.

REFERENCES

Cohen, Raviv and Ruths, Derek (2013) *Classifying Political Orientation on Twitter: It’s Not Easy!*, ICWSM.

Al Zamal, Faiyaz and Liu, Wendy and Ruths, Derek (2012) *Homophily and Latent Attribute Inference: Inferring Latent Attributes of Twitter Users from Neighbors*, ICWSM.

Mladenic, Dunja and Grobelnik, Marko (1999) *Feature selection for unbalanced class distribution and Naive Bayes*, ICML.

Users Location Prediction in Location-based Social Networks

Jarana Manotumruksa
University of Glasgow
j.manotumruksa.1@research.gla.ac.uk

The wealth of user-generated data in Location-Based Social Networks (LBSNs) has opened new opportunities for researchers to model and understand human mobile behaviour, including predicting where they are most likely to check-in next. In this paper, we propose a model that leverages the use of Global Temporal Preferences and Spatial Correlation, to help make predictions for a previously unseen user - the so-called cold-start problem. The experimental results on a real-world LBSN dataset show that our proposed model outperforms the state-of-the-art approaches on prediction accuracy and can alleviate the cold-start problem.

Location-Based Social Networks, User's Location Prediction, Cold-Start Problem

1. INTRODUCTION

Location-Based Social Networks (LBSNs), such as Foursquare and Brightkite, provide enormous user-generated data containing location data and human activity, in the form of *check-ins*, which can be exploited to understand the social and temporal characteristics of users on LBSNs. This includes predicting the user's location at a certain time, which can be useful for the design of future mobile location based services, traffic forecasting or urban planning.

In this paper, we address the issue of predicting the next location of an individual based on his/her historical check-in data. An existing state-of-the-art approach for user's location prediction proposed by Gao *et al.* (2013) used the user's daily and weekly cyclic check-in patterns to model his/her temporal preferences. However, LBSN check-in data is usually very sparse, resulting in difficulties when aiming to effectively model the personal temporal preferences of a user. To overcome this problem, Gao *et al.* (2013) proposed the use of smoothing techniques and also employ social correlation, i.e. using the preferences of the user's friends on the social network to improve suggestions.

Although the aforementioned approaches can effectively tackle the data sparsity problem, these approaches do not tackle the problem of previously unseen users, i.e. the cold-start problem. In this paper, we propose a model that leverage the use of Global Temporal Preferences and Spatial Correlation to alleviate the cold-start problem. Global Temporal Preferences exploits the historical check-ins of other users to model the temporal popularity of locations. Spatial Correlation estimates a distance that a user is willing to visit a location based on his/her current location.

The remainder of this paper is organised as follows. First, we review relevant related work in Section 2. We define the problem and our approach in Section 3. The experiment setup and results are described in Section 4. Finally, conclusions and direction for future work follow in Section 5

2. RELATED WORK

The availability of check-in data in LBSNs has recently attracted the researchers' attention. Gao *et al.* (2013) proposed to use the historical check-ins of users and their friends' in LBSNs to identify daily and weekly cyclic patterns in the users' mobile behaviour. In particular, they proposed to model the temporal preferences of a user with a Gaussian mixture model that estimates the distribution of user check-ins and predicts their location. In comparison with our proposed approach, we consider only the most recent check-in of a user as a user's current location and we use the historical check-in data of other users regardless of their relationship to the user, i.e. friendship, to model the popularity of locations at specific time. The most popular location at a specific time which is nearby the user current location is inferred as the user's next location.

Noulas *et al.* (2012) is the most related to our work where they considered the popularity of a location, i.e. the total number of check-ins at the location, as a feature and used supervised learning technique to predict the user's location. Deveaud *et al.* (2014) showed that the popularity of a location is an effective feature in a Learning-to-Rank technique for Point-of-Interest (POI) recommendation. In contrast to these works, we consider the temporal popularity of the location at a specific time period instead of the overall popularity.

Besides the popularity of a location, there have been some attempts to incorporate spatial influence for POI recommendation. Yuan *et al.* (2013) used a power law distribution to model the willingness of a user to visit a distant POI. In our work, we calculate an average distance between two successive check-ins based on their elapsed time. We then use these average distances as a threshold to filter out any locations that are far away from the most recent check-in location.

3. METHODOLOGY

In this section, we firstly explain the problem of predicting the user's location in LBSN in details (Section 3.1). We then describe our proposed model that consists of 2 components, Global Temporal Preferences (Section 3.2) and Spatial Correlation (Section 3.3).

3.1. Problem Definition

The problem of predicting the user's location in a LBSN can be formally defined as follows. Given a time t, the problem is to predict the location $l \in \mathcal{L}$ that user $u \in \mathcal{U}$ will visit based on his/her historical visits, $\mathcal{C}_{u,t}$, where $\mathcal{U}$ and $\mathcal{L}$ are the set of users and locations respectively and $\mathcal{C}_{u,t}$ is the set of check-ins for the user u before time t. Let $\mathcal{C}$ be the global set of all check-ins, with each check-in $c \in \mathcal{C}$ is represented as a tuple $\langle u, l, t \rangle \in \mathcal{C}$, indicating a check-in generated by a user $u \in \mathcal{U}$ at location $l \in \mathcal{L}$ at time t.

It is possible to represent a specific time t (e.g. "2015-02-15 17:45:22") as a *time-slot*, for instance as a specific hour of the day (17:00) or a day of the week (Sunday). Given a time t, $\mathcal{T}_m(t)$ is a function that returns a time slot w.r.t the specific time slot granularity m. For example, this function can be chosen to produce a time slot for each hour of the day, i.e. $\mathcal{T}_D(t) \in \{0, 1, \ldots, 23\}$.

Next, we describe how the global temporal preferences are modeled using the historical check-in data of all users.

3.2. Global Temporal Preferences

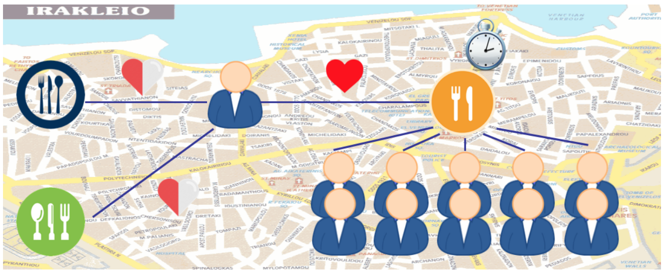


Figure 1: *An influence of mostly visited location to user's preferences*

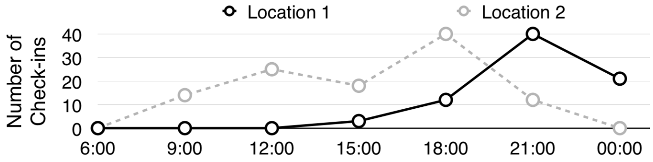


Figure 2: *A distribution of number of check-ins over time periods of two locations*

The popularity of a location, i.e. the total number of check-ins from all users on the location, is an important factor affecting human's check-in behaviour and has been exploited in venue recommendation and user's location prediction in earlier studies, e.g. Noulas *et al.* (2012); Deveaud *et al.* (2014). In this work, we assume that users are influenced by other users regardless of their relationships. Intuitively, as illustrated by Figure 1, if many users have visited a venue at a particular time, this venue may be more attractive to visit than other venues at that time. We can infer the popularity of a location l, as follows:

$$Popular(l) = | \{\langle u_i, l_j, t_k \rangle; \langle u_i, l_j, t_k \rangle \in \mathcal{C}_{u,t}, l_j = l\} | \quad (1)$$

Based on experimental check-in data from the Brightkite LBSN used by Gao *et al.* (2013), we found that the popularity of different locations varies over a time period as shown in Figure 2. Hence, to capture the temporal popularity of a location, we calculate the total number of check-ins of the location l at a specific time t, as follows:

$$Popular_m(l,t) = | \{\langle u_i, l_j, t_k \rangle; \langle u_i, l_j, t_k \rangle \in \mathcal{C}_{u,t}, l_j = l, \mathcal{T}_m(t) = \mathcal{T}_m(t_k)\} | \quad (2)$$

where m is the specific time slot granularity. To model the global temporal preferences of all users, we propose to compute the probability that the user will visit location l at a given time t, regardless of the user's historical check-in data, as follows:

$$P_m(l,t) = \frac{Popular_m(l,t)}{Popular(t)} \quad (3)$$

3.3. Spatial Correlation

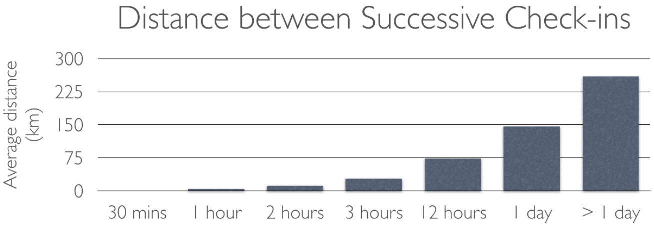


Figure 3: *Distance between successive check-ins*

In the previous section, we described how to predict a probability that a user will check in at location l by the global temporal preferences using Equation (3). However, capturing the user mobile behaviour in LBSNs solely using Global Temporal Preferences is insufficient. Yuan *et al.* (2013) suggested that users are more willing to check-in at nearby locations to their current location.

In our initial analysis using an experimental Brightkite dataset used by Gao *et al.* (2013), we found that there is a correlation between successive check-ins. Figure 3 shows a distribution of the distance between successive check-ins over time periods. In particular, the distance between two check-ins is correlated with the elapsed time of these two check-ins. Namely, within a short period of time, users are more likely to check-in at a location close to their previous check-in.

We calculate the elapsed time between the testing check-in and the most recent check-in. Then we filter out those locations whose distance to the most recent check-in location is larger than a threshold, namely an average distance with respect to the elapsed time. The qualified locations will be ranked based on their distance to the location of the user's most recent check-in using the following equation:

$$P_d(l) = dist(l, l_{recent}) \tag{4}$$

where $dist$ is a function that returns a distance between two locations in kilometers which are calculated using the Haversine Formula Shumaker, B. P., and R. W. Sinnott (1984). We use a linear combination to incorporate Spatial Correlation (Equation (4)) with Global Temporal Preferences (Equation (3)) as follows:

$$P_m(l,t) = \alpha P_d(l) + (1-\alpha) P_m(l,t) \tag{5}$$

where α is a parameter that controls the relative contributions of Global Temporal Preferences and Spatial Correlation.

4. EXPERIMENTS AND RESULTS

In this section, we report experiments conducted to evaluate the effectiveness of our approach in alleviating the cold-start problem using a real-world check-in data from a LBSN.

Table 1: *Salient statistics of the Brightkite LBSN dataset.*

Duration	04/2008-10/2010
# of Users	26,915
# of Check-ins	4,532,151
# of Unique Locations	751,176
# of Test Check-ins	134,575
Average Check-ins per user	168

Dataset. The publicly available check-in data from Brightkite[1] is used in our experiment. Salient statistics of the dataset are listed in Table 1.

Setup. We consider 26,915 users who have at least 10 check-ins in evaluating the effectiveness of our approach (All Users). We also consider 8,613 users who have less than 20 check-ins in evaluating the extent to which our approach alleviates the cold-start problem (Cold-Start Users). Both experiments are conducted using a 5-fold cross validation on a user level, where for each testing user, we randomly select 5 check-ins as test check-ins. For each test check-in, we consider its check-in time t as given, its check-in location as the ground truth data, and a set of check-ins of a user before time t, $\mathcal{C}_{u,t}$, as observed data. We rank all locations extracted from the observed data based on their prediction scores using the Spatial Correlation and Global Temporal Preference model (SGTP) in Equation (5). Then we select the top ranked location as the predicted location where the user is most likely to visit next. To set α, we use 5-fold cross validation, varying α from 0.0 to 1.0 in 0.1 incremental steps, to determine the value that maxmimises Success@1 (see below).

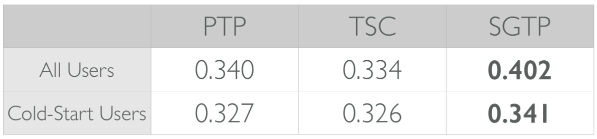
Measure We evaluate the accuracy of the predicted locations using Success@1, which was called *prediction accuracy* by Gao *et al.* (2013), i.e. the ratio of the number of accurately predicted locations to the total number of testing check-ins.

Baselines Two baselines used in our experiments are the state-of-the-art user's location prediction approaches proposed by Gao *et al.* (2013) : (i) Personal Temporal Preference (PTP) where they predict the next location based on the user's daily and weekly historical check-in data and (ii) Temporal Social Correlation (TSC) where they use a collaborative filtering technique to predict the next location based on the temporal preferences of user's friends.

Table 2 shows the prediction accuracy of our approach (SGTP) in comparison with the baselines

[1] http://snap.stanford.edu/data/loc-brightkite.html

Table 2: *Prediction accuracy (success@1) for various models*



	PTP	TSC	SGTP
All Users	0.340	0.334	**0.402**
Cold-Start Users	0.327	0.326	**0.341**

(PTP and TSC). The first row of the table presents the results of the effectiveness evaluation using 26,915 users. The second row of the table compares the effectiveness in alleviating the cold-start problems between our approach and the baselines using 8,613 users. From Table 2, we observe that the linear combination of Spatial Correlation and Global Temporal Preference (SGTP) improves the prediction accuracy by 18% and 20% in comparison with Personal Temporal Preferences (PTP) and Temporal Social Correlation (TSC), respectively. SGTP obtains the optimal results when α is set to 0.9. This clearly demonstrates that users are more likely to check-in at nearby locations than at the most popular ones. In particular, the spatial correlation plays a more important role in predicting the user's location than the global temporal preferences in this dataset. Finally, our approach is promising in alleviating the cold-start problem more effectively than the baselines, by approximately 4% (0.341 vs 0.327 and 0.326, respectively).

5. CONCLUSIONS AND FUTURE DIRECTIONS

The availability of historical check-in data in LBSNs can be exploited to understand the user mobile behaviour. In this paper, we propose a model that leverages Global Temporal Preferences and Spatial Correlation to alleviate the problem of unseen users (cold-start users). The experiment results on real-world check-in data in a LBSN shows that our proposed approach outperforms the stat-of-the-art approaches both in terms of effectiveness and in alleviating the cold-start problem.

The study of user mobile behaviour on LBSNs can be exploited in many applications. A venue recommendation or geo-based advertising application could take this into account in order to improve its effectiveness based on the user's location and the time of the day.

6. ACKNOWLEDGMENTS

I would like to thank my supervisors, M-Dyaa Albakour, Craig Macdonald and Iadh Ounis for their supports. We thank ACM SIGIR for the awarded scholarship for participating in ESSIR 2015 and the FDIA Symposium.

REFERENCES

Gao, Huiji and Tang, Jiliang and Hu, Xia and Liu, Huan (2013) *Modeling temporal effects of human mobile behavior on location-based social networks. Proceedings of CIKM, pp. 1673-1678*.

Noulas, Anastasios and Scellato, Salvatore and Lathia, Neal and Mascolo, Cecilia (2012) *Mining user mobility features for next place prediction in location-based services. Proceedings of ICDM, pp. 1038-1043*

Deveaud, Romain and Albakour, Dyaa and Macdonald, Craig and Ounis, Iadh and others(2014) *On the Importance of Venue-Dependent Features for Learning to Rank Contextual Suggestions. Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, pp. 1827-1830*

Deveaud, Romain and Albakour, Dyaa and Macdonald, Craig and Ounis, Iadh and others(2014) *Time-aware point-of-interest recommendation. Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval, pp. 363-372*

Shumaker, B. P., and R. W. Sinnott (1984) *Astronomical computing: 1. Computing under the open sky. 2. Virtues of the haversine." Sky and telescope 68, pp. 158-159*

Sentiment analysis via fractal dimension

Symeonidis Symeon
Department of Electrical and Computer Engineering
Democritus University of Thrace, Xanthi 67100, Greece
ssymeoni@ee.duth.gr

This paper presents a very early-stage idea about sentiment analysis with usage of theory of fractal dimension. It presents a short literature review about fractal dimension on datasets and an approach to sentiment analysis with fractal dimension.

Sentiment analysis, Fractal Dimension, Classification

1. INTRODUCTION

Everyday more and more people use the Internet in order to communicate,work, search information or fro personal. The existing knowledge in conjunction with the vast amount of information collected by the use of Internet and the new ideas generated, imposed a shift in the scientists’ to a more customize science which is callled opinion mining and sentiment analysis.

Many applications for opinion mining and sentiment analysis have been deployed to analyze opinions, feelings and attitudes (Maks and Vossen 2012). In their previous studies, various authors categorized sentiment in three polars (da Silva, Hruschka, and Hruschka 2014) or in six “universal” emotions (Lemeignan, Guitart, and Bloch 1991). However is it possible for the dimensions of emotions to be predicted and classified? Furthermore, can a feeling be positive or negative and to what extent?

Fractal dimension is defined as “an estimate of the degrees of freedom of a data set” (Kumaraswamy 2003). Fractal analysis is employed by various sciences such as education and medicine, with main purpose to better quantify and describe, the deviation and complexity of images using a computable value (Park, Wang, and Zheng 2009). Despite the several limitations in providing accurate measures, fractal dimension and Euclidean geometry have provided crucial solutions to complicated everyday problems (Florindo and Bruno 2014).

An interesting approach, is this of information retrieval and specific of opinion retrieval - mining and sentiment analysis of users with the use of fractal dimension. The literature gap in this research area is a challenge in order to better explore and approach this issue.

2. RESEARCH DESCRIPTION

The last couple of years a high research interest is observed in the field of multimedia database management systems and more precisely in this of multimedia processing within computing systems (Nappi, Polese, and Tortora 1998). Application of fractal dimension cannot numerate sharply any objects and for such these objects is computed by estimation or approximation (Sadikin & Ito 2013).

Many data mining tasks such as dimensionality reduction, classification, clustering, learning patterns are used in sentiment analysis, mainly as patterns and an indicator of the way the data points are spread in the data space(Kumaraswamy 2003). In addition, the relationship between the spread of the data and the amount of information that can obtained, can enhance the performance of a given data mining method which is evaluated on the basis of the information captured. Nevertheless, all the above mentioned are expensive and require a large computation time during their implementation (Kumaraswamy 2003).

2.1 Data Fractal Dimension

As a dataset has fractal character, opinion and sentiment data can have the same capacity, if their properties such as the structure and the statistic distribution are the same in partial distribution(Ni, Ni, and Gao 2011; Bao et al. 2004).

Yan et al. (2006) measured the fractal dimension Dq of dataset by:

$$D_q = \begin{cases} \lim\limits_{r \to 0} \frac{\sum_i p_i \log p_i^q}{\log r} & q = 1 \\ \lim\limits_{r \to 0} \frac{1}{q-1} \frac{\log \sum_i p_i^q}{\log r} & q \neq 1 \end{cases} \quad r \in [r_1, r_2]$$

Figure 1: *Fractal dimension Dq of dataset*

(c) Symeonidis Symeon. Published by BCS
Learning and Development Ltd.
Proceedings of the 6th Symposium on
Future Directions in Information Access 2015

where r is the grid size, pi is the probability of the data points in the ith grid, q is an integer (Ni, Ni, and Gao 2011).

2.2 Fractal Information Retrieval (FIR)

According to (Traina et al. 2010) “fractal dimension of an Euclidean object corresponds to its Euclidean dimension and it is always an integer number”. Moreover Zhang et. al (2002) explain a technique which can discover and select the number of significant attributes to describe a dataset using fractal dimension. The study of combining Information Retrieval and fractal dimension can importantly enhance the results of retrieval techniques.

2.3 Fractal Sentiment Analysis (FSA)

Different techniques of sentiment analysis were published in the last years. Some of these approaches are based on natural language processing, lexicons and machine learning.

The problem in analyzing sentiment is how to convert every word to fractal. Is it necessary for every word to be converted according to the part of speech that it is (articles, nouns, pronouns, adjectives, verbs, adverbs, conjunctions, prepositions, and interjections)? Which technique of sentiment analysis is more appropriate when analyzing sentiment via fractal dimension?

My first approach is based on the ground that every word is related to the fractal dimension of a sentence and every word is correlated to the previous or following one. Every correlated word increases the complexity but the dimensional reduction method can avoid redundant dimensions and recover the original variables while preserving the topological properties of the sentence (Kumaraswamy 2003). Furthermore, in order to measure the spread of words and the intrinsic dimension of the sentence, fractal dimension is used which calculates the correlation integral and classification errors.

Another approach is this of computing every word with sentiment fractal dimension according to the part of speech that this is and by using priority to parts with “specific and essential” sentiment. Following the previous procedures, can be classified and different fractal levels for every category will be computing. This approach must be based on a sentiment analysis technique with strong sentiment evaluation for every word.

3. CONCLUSIONS

A very early-stage idea about fractal sentiment analysis was presented above. Further research in the existing literature combined with personal ideas and assumption s, will probably lead us to results.

4. ACKNOWLEDGEMENT

I would like to thank ELIAS (Evaluating Information Access Systems) ESF Research Networking Programme for the awarded scholarship for participating in ESSIR 2015 and the FDIA Symposium. Also I would like to thank my supervisor Avi Arampartzis.

5. REFERENCES.

Bao, Yubin, Ge Yu, Huanliang Sun, and Daling Wang. 2004. *Advances in Web-Age Information Management*. Edited by Qing Li, Guoren Wang, and Ling Feng. *Advances in Web-Age Information Management: 5th International Conference, WAIM 2004, Dalian, China, July 15-17, 2004*. Vol. 3129. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/b98703.

Da Silva, Nádia F F, Eduardo R. Hruschka, and Estevam R. Hruschka. 2014. “Tweet Sentiment Analysis with Classifier Ensembles.” *Decision Support Systems* 66. Elsevier B.V.: 170–79. doi:10.1016/j.dss.2014.07.003.

Florindo, João Batista, and Odemir Martinez Bruno. 2014. “Fractal Descriptors Based on the Probability Dimension: A Texture Analysis and Classification Approach.” *Pattern Recognition Letters* 42 (1): 107–14. doi:10.1016/j.patrec.2014.01.009.

Kumaraswamy, Sree Krishna. 2003. “Fractal Dimension for Data Mining.” *Ml.Cmu.Edu*. http://www.ml.cmu.edu/research/dap-papers/skkumar_kdd_project.pdf.

Lemeignan, M., P.L. Guitart, and S. Bloch. 1991. “Autonomic Differentiation of Six Basic Emotions.” *International Journal of Psychophysiology* 11 (1): 52–53. doi:10.1016/0167-8760(91)90222-J.

Maks, Isa, and Piek Vossen. 2012. “A Lexicon Model for Deep Sentiment Analysis and Opinion Mining Applications.” *Decision Support Systems* 53 (4). Elsevier B.V.: 680–88. doi:10.1016/j.dss.2012.05.025.

Nappi, M., G. Polese, and G. Tortora. 1998. “FIRST: Fractal Indexing and Retrieval SysTem for Image Databases.” *Image and Vision Computing* 16 (14): 1019–31. doi:10.1016/S0262-8856(98)00084-5.

Ni, Li-Ping, Zhi-Wei Ni, and Ya-Zhuo Gao. 2011. “Stock Trend Prediction Based on Fractal Feature Selection and Support Vector Machine.” *Expert Systems with Applications* 38 (5). Elsevier Ltd: 5569–76. doi:10.1016/j.eswa.2010.10.079.

Park, Sang Cheol, Xiao-Hui Wang, and Bin Zheng. 2009. “Assessment of Performance Improvement in Content-Based Medical Image Retrieval Schemes Using Fractal Dimension.” *Academic Radiology* 16 (10): 1171–78. doi:10.1016/j.acra.2009.04.009.

Sadikin, Mujiono, and Wasito Ito. n.d. “FRACTAL DIMENSION AS A DATA DIMENSIONALITY REDUCTION.” *The 7th International Conference on Information & Communication Technology and Systems (ICTS) 2013, 15 – 16 May 2013, Surabaya*, 105–10.

Traina, C, Agma Traina, Leejay Wu, and Christos Faloutsos. 2010. “Fast Feature Selection Using Fractal Dimension.” *Journal of Information and Data Management* 1 (1): 3–16. http://repository.cmu.edu/compsci/580/.

Yan, Guanghui, Zhanhuai Li, and Liu Yuan. 2006. *MICAI 2006: Advances in Artificial Intelligence*. Edited by Alexander Gelbukh and Carlos Alberto Reyes-Garcia. *MICAI 2006: Advances in Artificial Intelligence, 5th Mexican International Conference on Artificial Intelligence, Apizaco, Mexico, November 13-17, 2006, Proceedings*. Vol. 4293. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/11925231.

Zhang, Haiqin, Chang-Shing Perng, and Qingsheng Cai. 2002. “An Improved Algorithm for Feature Selection Using Fractal Dimension.” *Proceedings of the Second International Workshop on Databases, Documents, and Information Fusion*. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.87.4407&rep=rep1&type=pdf.

Investigating Search Behavior and Performance using Personal and Social Context Signals

Dongho Choi
Rutgers, The State University of New Jersey
4 Huntington Street
New Brunswick, New Jersey 08901
United States
dongho.j.choi@rutgers.edu

Emerging trends in smart-phones and wearable devices provide us efficient way through we could observe and understand individuals' behavior and allow for the creation of a rich user behavioral profile. Through a user study, personal and contextual signals from participants' everyday lives were collected, while information search behavior was observed in a lab study. Preliminary analysis over data indicates that possibilities of relationship between social/geo-locational interaction between information search performance.

search behavior, social context, wearable devices

1. INTRODUCTION

Emerging technologies and relating devices help people find information and share their knowledge and experience with ease. In addition to the traditional desktop PC environment and smart-phones, diverse types of wearable devices are also quickly emerging as information devices/systems through which people search, produce, and share information. While smart-phones make easier ways to quickly seek information and communicate people, wearable devices tend to be focusing on personalized information, such as notification for pre-defined request and users' biophysical signals. More specifically, various types of sensor on these devices help us understand people's behavior in everyday lives from diverse aspects. Personal and social context information has become much approachable for researchers to study both individual and social behavior in details thanks to the devices and sensors. Interested in exploring signals from various types of sensors and/or devices, a user study was conducted aiming to investigate individual and collaborative search and the factors that relate to them.

2. BACKGROUND

An individual, the person who consumes and produces information, has several intrinsic characteristics, such as general behavioral style and personality, as well as preferences toward social relationship and physical activities. Multiple studies have acknowledged the effect of individual personality, demographic descriptors, and personal context in human information behavior (Burdick 1996). However, most studies had to focus on traits that could be easily observed (e.g. gender, ethnicity) or elicited in a short time in laboratory settings. In effect the human personal context was captured by observations made in short, unnatural settings, as recorded or reported by participants or the study conductor and had to contend with multiple challenges, including subjectivity in observation, recall/cognitive/socio-cognitive biases, and limited observation opportunities.

Meanwhile, collaboration is considered an activity where individual characteristics and social context work simultaneously in complicated manners. Multiple number of person who has different characteristics and background gather and work together to achieve common goal(s). And the basic assumption and/or expectation of collaboration is that through working together physically or timely, they obtain

Proceedings of the 6th Symposium on
Future Directions in Information Access 2015

improved outcomes both in quantitatively and qualitatively. Collaboration often involves working with information, such as collaborative search. Collaboratively searching for information can be find in many domains ranging from education (Hyldegård 2006) to health-care (Reddy and Jansen 2008). However, we lack of understanding when and how people working together could take the advantage of collaboration and the fruits; what kinds of combination of people benefits collaboration regarding their unique characteristics and contexts.

The emergence of mobile phones and sensors that people wear while undertaking activities of daily living is allowing researchers to create rich personalized models of human behavior in social, spatial, and temporal contexts and connect them to various types of outcomes. Singh et al. (2013) used social behavior that was measured through face-to-face interaction, phone call, and SMS use logs, to predict the spending behavior regarding visiting to diverse businesses and overspending. Mobile phone sensors are also used to develop context-aware systems that detects when individual is depressed and he/she requires assistance (Burns et al. 2011), and emotion sensing platform (Rachuri et al. 2010; Yano et al. 2012).

The research questions I have are as follows:

RQ1. To what extent, if any, do behaviors observable via mobile phones relate to information seeking behaviors when people work in individual and/or collaboration?

RQ2. To what extent, if any, do behaviors observable via wearable devices relate to collaborative search behavior?

RQ3. To what extent, if any, are personal signals and social signals different in terms of predicting the search performance?

3. STUDY

3.1. Session Flow

25 student participated in this study through three batches of session during Spring semester in 2015. The workflow of each session is described in Table 1. The participants were invited to a lab for the introduction during which required apps were installed on their phones and a wearable device was given to each of them. During the field study, participants were not asked to do anything specific except keeping the mobile app installed and wearing the device on their wrist on daily life with regular (automatic, in the background) syncing to the dedicated app.

3.2. Lab session

During the lab session, participants were asked to collect relevant information in an exploratory search task on the given topics. The first task (Task 1) was done individually with the topic of "data security, " while the second task (Task 2) was done in pair randomly assigned with the topic of "health and wellness".

3.3. Evaluation

In order to measure and evaluate the participants' information searching behaviors, and compare the differences between those from individual tasks and collaborative tasks, the study adopted a subsets of evaluation measures suggested in Shah (2014) and Shah et al. (2015). A brief description of the framework is presented in Table 2.

4. RESULTS AND DISCUSSION

In this session, results of preliminary correlation analysis between the variables in Table 2 are presented.

4.1. Individual Search Performance

Regarding data collected through mobile app, two features show correlation with some of search performance. The *number of SMS* has positive relationship with *Unique Relevant Coverage*, while *Distinct Location* also has positive relationship with *Unique Coverage*.

The variable of *Number of SMS* shows the extent to which a person interacts with other individuals through mobile texts. Higher value of this variable means a person frequently sends and receives text messages to and from other people no matter who they are. In the meantime, higher value of *Unique Relevant Coverage* indicates that a person visited relatively much number of relevant Web pages during the search task. More specifically, when considering the term "unique," the person might have distinct and unique criteria for relevancy. When it comes to the exploratory search, the tendency also infers that an individual who sends/receives relatively higher number of messages per day tends to understand and learn the topic differently from others who does not.

One additional interesting finding is the positive relation between *Unique Coverage* and *Distinct Location*. The variable of *Distinct Location* measures that the total number of distinct locations during a day, captured by different cell tower IDs and/or GPS data. This information does not necessarily imply the intention of locational changes, whether

Table 1: *Session Workflow.*

Session	Procedure	Description
Field Session	Introduction	Introduce the study, install required apps and sync the wearable device with a dedicated app.
	Field task	Have participants keep using wearable devices in everyday lives and apps on the phone, collecting their individual and social context data.
	Survey	In the middle of the field session period, ask about their behavioral style, social capital, and information behavior.
Lab Session	Introduction	Introduce the lab session and information-seeking tasks the participants will be given.
	Lab task 1	Exploratory search task-1. Includes pre-survey and post-survey.
	Lab task 2	Exploratory search task-2. Includes pre-survey and post-survey.
	Wrap-up	Wrap-up study and (optional) interview.

for activities or commuting. No matter what the purposes of movement, the result indicate that a person who moves more tends to visit Web pages that have not visited by other participants. This brings us an interesting question: “Does geo-locational movement affect information exposure to an individual and affect again toward the way of thinking and understanding?”

4.2. Collaborative Search Performance

Among the features extracted signals, we found relationship between the dissimilarity in the number of SMS counterparts and the synergy effect. This means that the more participants are different with regard to the social activities that are measured how many people they are keeping to talk with, the less synergy effect they have when working together. Though it is not significant enough, we can see the negative relationship between distinct users for SMS and the ratio of increased *Coverage*. However, if we selected sample with negative synergy effect, the relationship is much stronger and significant.

An another aspect of SMS usage, the number of SMS that an individual sends and receives per day, shows a relationship with the synergy effect on *Relevant Coverage* in collaboration ($N = 12, R^2 = .344, p = .045$). This implies that if members in a pair have substantial difference of the extent of using SMS, they tend to have less synergy effect regarding information relevancy.

Seeing the data, even with very limited number of samples, we came up with an interesting question: “Do the large differences in the values of the same feature from the collaborators matter? Or does the minimum value of each feature among the collaborators matter more for group synergy?” This question is inspired by the “Liebig’s Law of Minimum (Odum et al. 1971),” which is originally a concept applied to plant or crop growth. It was found that increasing the amount of plentiful nutrients did not cause the increasing plat growth and the yield. Likewise, we suspected even if one participant in a team has incredible potential in terms of productivity and efficiency in the task, the results might not be productive as expected, when the other one cannot support him/her very well.

4.3. Limitations and Future Work

There are some limitations on this study. For example, we need to consider the changing techniques and applications people use for chatting, instead of using the conventional simple text messaging. We can easily name a couple of popular chatting applications for smart-phones, which have been replacing texts. The way of using those apps maybe a lot different from the collected data in this study. If we could use a method through which observing interpersonal communication over smart-phones, we can conduct much different kinds of experiments and analysis.

Given that only limited number of features were used, we want to see more different aspects of the collected data. We are expecting to see more diverse perspectives toward the data through visualization and deep analysis with additional aspects, such as diversity, loyalty, regularity and so on.

5. CONCLUSION

Personal and social contextual signals were collected through smart-phone app and wearable device, while information search behavior was monitored during exploratory search tasks. Preliminary analysis over the extracted features from the raw data indicates several interesting points. Regarding individual task, *the number of SMS* has positive relationship with *Unique Relevant Coverage*, and *Distinct Location* has a positive relationship with *Unique Coverage*. When it comes to the collaborative task, social interaction (SMS usage) via phones and physical activities have relation with the synergy

Table 2: *Variables used in Analysis*

Data Source	Features	Description
Wearable Device	Steps	The average number of steps per day of a user during the field study
	Minutes of Sedentary	Average minutes of being sedentary per day recognized by the wearable device
	Minutes of Very Active	Average minutes of being very active per day recognized by the wearable device
	Minutes of Sleep	Average minutes of being asleep per day recognized by the wearable device
Mobile App	Total Call Time	Average minutes of total call time per day
	Number of Calls	Average number of call per day
	Number of SMS	Average number of SMS sent/received per day
	Distinct User (Call)	Average number of distinct caller/receiver per day
	Distinct User (SMS)	Average number of distinct sender/receiver for SMS usage per day
	Distinct Location	Average number of distinct location per day, captured by the app using the number of cell tower around and GPS data
Searching Task	Coverage	The total number of distinct Web documents visited by a participant
	Unique Coverage	The total number of Web documents visited only by one participant
	Relevant Coverage	The total number of distinct relevant Web documents visited by a participant
	Unique Relevant Coverage	The total number of distinct relevant Web documents visited only by one participant
	Distinct Queries	The number of distinct queries that were submitted to search engines by participant

effect of collaboration. Also, our results implies that the difference between collaborators matter with regard to social interaction, while minimum extent of physical activity relate to the outcomes.

ACKNOWLEDGEMENT

I thank ELIAS (Evaluating Information Access Systems) ESF Research Networking Programme for the awarded scholarship for participating in ESSIR 2015 and the FDIA Symposium.

REFERENCES

Burdick, T. A. (1996), ‘Success and Diversity in Information Seeking: Gender and the Information Search Styles Model.’, *School Library Media Quarterly* **25**(1), 19–26.

Burns, M. N., Begale, M., Duffecy, J., Gergle, D., Karr, C. J., Giangrande, E. and Mohr, D. C. (2011), ‘Harnessing context sensing to develop a mobile intervention for depression.’, *Journal of medical Internet research* **13**(3), e55.

Hyldegård, J. (2006), ‘Collaborative information behaviour—-exploring kuhlthaus information search process model in a group-based educational setting’, *Information Processing & Management* **42**(1), 276–298.

Odum, E. P., Odum, H. T. and Andrews, J. (1971), *Fundamentals of ecology*, Vol. 3, Saunders Philadelphia.

Rachuri, K. K., Musolesi, M., Mascolo, C., Rentfrow, P. J., Longworth, C. and Aucinas, A. (2010), ‘EmotionSense’, *Proceedings of the 12th ACM international conference on Ubiquitous computing - Ubicomp ’10* p. 281.

Reddy, M. C. and Jansen, B. J. (2008), ‘A model for understanding collaborative information behavior in context: A study of two healthcare teams’, *Information Processing & Management* **44**(1), 256–273.

Shah, C. (2014), ‘Evaluating collaborative information seeking - synthesis, suggestions, and structure’, *Journal of Information Science* **40**(4), 460–475.

Shah, C., Hendahewa, C. and González-ibá nez, R. (2015), ‘Two’s Company, But Threes No Crowd: Evaluating Exploratory Web Search for Individuals and Teams’, *Journal of Information Science* **Sep**.

Singh, V., Freeman, L., Lepri, B. and Pentland, A. (2013), ‘Classifying spending behavior using socio-mobile data’, *HUMAN* **2**(2), pp–99.

Yano, K., Lyubomirsky, S. and Chancellor, J. (2012), ‘Sensing happiness’, *IEEE Spectrum* **49**(12), 32–37.

Word-Context Matrix based Query Expansion in Information Retrieval for Turkish Text

Emre Satir
Department of Computer Engineering
Dokuz Eylul University, Izmir, TURKEY
emre.satir@st.cs.deu.edu.tr

Adil Alpkocak
Department of Computer Engineering
Dokuz Eylul University, Izmir, TURKEY
alpkocak@cs.deu.edu.tr

Deniz Kilinc
Faculty of Technology
Celal Bayar University, Manisa, TURKEY
deniz.kilinc@cbu.edu.tr

In this paper, we proposed a Query Expansion (QE) approach on a Turkish Text collection based on word-context matrix with a sliding fixed sized window and Singular Value Decomposition (SVD) method. Our query expansion approach uses semantic relationship of terms to improve the existing query expansion methods available in the literature, namely Bo1 (Bose-Einstein 1), Bo2 (Bose-Einstein 2) and KL (Kullback-Leibler). We evaluated our approach on Milliyet collection, which is a Turkish IR test bed containing more than 400K documents and 72 queries. The experimentation shows that our approach clearly improves the all three QE methods in terms of major Information Retrieval (IR) performance measures such as MAP, R-precision and P@10.

Word-context matrix, SVD, Query expansion, Turkish information retrieval

1. INTRODUCTION

As a result of rapid development of Internet and related technologies, the amount of data is growing day by day. This data has to be organized and retrieved whenever it is needed. Information Retrieval (IR) is a discipline that is related with indexing, retrieving, and structuring documents from any collections. The retrieval methods aim to find the best matching documents according to user query (user need) within a large document collection. In general users may not know how to construct the best query according to their needs and the queries may be inadequate. Query Expansion (QE) is the process of reformulating the basic user query in order to get a better retrieving performance. Different techniques can be conducted to expand a query such as using synonyms of terms, using ontologies to add related terms, or checking spelling errors of terms and correcting them.

In this paper, we proposed a word-context matrix (Turney & Pantel, 2010) based QE on Turkish Text. Although we have applied our technique to Turkish Text, it is applicable to all languages (i.e. it is language independent). We tried to find expanded terms not outside of the collection (like using an ontology) but from the collection itself. First we constructed a huge term co-occurrence matrix from the text collection and applied Singular Value Decomposition (SVD) method in order to obtain a reduced word-context matrix. Then we utilized this matrix to expand queries.

When applied to document similarity, SVD is called Latent Semantic Indexing (LSI), and when applied to word similarity, it's called Latent Semantic Analysis (LSA). (Turney & Pantel, 2010). Researchers, in their study (Landauer & Dumais, 1997) applied SVD to word similarity with using a term-document matrix. But our computations are based on a word co-occurrence matrix with a fixed-sized sliding window. We used these similarity computations to expand queries.

The rest of this paper is organized as follows. In section 2, our proposed method is introduced. Section 3 describes the experimental study, section 4 discusses the experimental results obtained and finally section 5 tells about conclusions and future work.

2. PROPOSED METHOD

The distributional hypothesis in linguistics is that words that occur in similar contexts tend to have similar meanings (Harris, 1954). A word context can be represented with context matrix. In general, in a word–context matrix, the context is given by different context such as blindly separated set of word window, or more morphologically by sentences, paragraphs, chapters, and documents. This context can be an extension for Vector Space Model (VSM) to measuring word similarity. A word may be represented by a vector in which the elements are derived from the occurrences of the word in various contexts. Then, similar row vectors

Proceedings of the 6th Symposium on Future Directions in Information Access 2015

in the word context matrix indicate similar word meanings.

In this study, we firstly constructed a word-context matrix by using S-Space Package that is an open source framework for developing and evaluating word space algorithms. We selected a fixed window size and use a sliding window style in order to count word co-occurrences. In order to compute word-similarities, we performed comparisons between their co-occurrence vectors. There are different types of similarity measures in the literature, herein; we used well-known cosine similarity measure.

If dimensionality of the matrix is reduced before computing semantic similarities, results can be improved (Landauer & Dumais, 1997). A good mathematical way of realizing this is using Singular Value Decomposition (SVD) (Landauer & Dumais, 1997). Actually SVD is a dimension reduction technique that allows to significantly reduce the number of columns so smaller matrix has the advantage that all subsequent similarity computations are much faster.

After creating our word-context matrix, we utilized SVD operation on this matrix for dimensionality reduction. For the first experiment, we used only this matrix to expand queries, and for the second experiment we used a combined method with the QE mechanisms implemented in Terrier IR platform.

For the second experiment (hybrid system), the procedure that we follow is:

- First we run Terrier with the one of the weighting models (Bo1, Bo2 or KL) and got the expanded terms.
- Secondly for every word in the basic query, we run our word co-occurrence based system with the cosine threshold 0.9 (actually we use distance not similarity, i.e. one minus the cosine of the included angle between points). We choose 0.9 because we want to achieve most of the words that close to the seed word. We discard only words that are very far away from the seed word.
- Then look for every word that Terrier found for the expanded terms in our word pool. If a word is included in both results, we added this word to the query (i.e. we filter out some words from QE mechanism implemented in Terrier).
- Finally we re-run Terrier with the original query and filter-outed expanded terms. We use the same weights taken from methods implemented in Terrier for the terms.

3. EXPERIMENTAL STUDY

In this study, our aim is to generate a new query expansion method based on word-context matrix and SVD, and to evaluate it experimentally. We used The Terrier IR Platform that is developed at the School of Computing Science, University of Glasgow. It is open source and written in Java. It is efficient and effective search engine which implements indexing and retrieving operations for large-scale collection of documents.

3.1 Data Set

In the study, we used Bilkent Milliyet Collection (Can, et al., 2008) that contains 408,305 documents (news articles and columns of five years, 2001 to 2005) from Turkish newspaper Milliyet. Each document includes approximately 234 words on the average. The collection contains about 95.5 million words. There are also 72 ad-hoc queries that are evaluated by 33 assessors. The query file includes Topic, Description and Narrative fields, but we only use Topic field in this study named as short query at (Can, et al., 2008). Table 1 presents first 5 short queries and their English translations.

Table 1: *First 5 short queries*

Query No	Topic (in Turkish)	Topic (English translation)
1	*Kuş Gribi*	*Bird Flu*
2	*Kıbrıs Sorunu*	*Cyprus Issue*
3	*Üniversiteye giriş sınavı*	*The university entrance exam*
4	*Tsunami*	*Tsunami*
5	*Mavi Akım Doğalgaz Projesi*	*Blue Stream Natural Gas Project*

3.2 Pre-processing and Indexing

Pre-processing is an important step before indexing. Tokenization, stop-word elimination and stemming are the most widely used pre-processing methods. In this study, we pre-processed the related tag contents. First, we converted all uppercase letters to lowercase equivalents and then we converted Turkish special characters to their Latin alphabet counterparts (ç→c, ğ→g, ı→i, ö → o, ş→s , ü→u). Also the same two operations are applied to the all fields of the query file. For stemming, we used fixed prefix stemming (Can, et al., 2008) and we selected the first 5 characters of a term as its stem. We employed a semi-automatically generated stop-word list contains 147 words that taken from (Can, et al., 2008). After pre-processing step, we indexed headline and text fields of the documents by using Terrier. We performed TF×IDF weighting model (Manning, et al., 2009) for our experiments.

3.3 Preparing the word co-occurrence matrix

As mentioned before we used S-Space package to construct the matrix. Before creating the matrix, we made some operations on Milliyet collection. In brief, stop-words removal operation is performed headline and text fields contents, the first 5 characters of a term is selected as its stem and Turkish special characters are converted to their Latin alphabet counterparts.

After these operations, we processed the converted collection with the S-Space package system. Total vocabulary (unique term size) is 139,916. For window size we used 5. SVD is a computationally exhausting operation so we decided not to use all terms in the vocabulary. We set the minimum frequency to 20, so we restricted our vocabulary to all terms occurring at least 20 times in the collection. So our new vocabulary size is 34,839 which led to a matrix of size 34,839x34,839. Finally we made SVD operation on this matrix with the reduction parameter of 400. So this gave us final matrix of size 34,839x400.

4. EXPERIMENTAL RESULTS

We utilized trec_eval IR evaluation tool for evaluation purposes. First we performed our baseline evaluations without any query expansion. Baseline results can be seen from the Table 2.

4.1 QE Methods Implemented in Terrier Results

Pseudo relevance feedback (Blind relevance feedback) uses a method like automatic local analysis. (Manning, et al., 2009). The user doesn't give feedback to the system, because the system automatically reformulates the query. First the system does the normal retrieval according to initial query and after finding documents the system assumes that the top *k* ranked documents are relevant, and finally to do relevance feedback under this assumption. Generally the procedure is:

- Take *k* documents from the initial retrieving as relevant results.
- Select top *t* terms from these documents using for example TF×IDF weights.
- Add these terms to query, do the retrieving process again and finally return new results.

Terrier has a built-in query expansion functionality that utilizes pseudo relevance feedback. In this study, we experimented with three query expansion models of Terrier: Bo1 (Bose-Einstein 1), Bo2 (Bose-Einstein 2) and KL (Kullback-Leibler). (Amati, 2003) Each model used for expanding the query with the most informative terms of the top-ranked documents. We experimented with only default values, which are 10 for the number of terms to expand a query with, and 3 for the number of top-ranked documents from which these terms are extracted.

Terrier utilizes a particular Divergence From Randomness (DFR) term weighting model to weight terms in the top-returned documents Bo1 model uses the Bose-Einstein statistics (Amati, 2003):

$$w(t) = tf_x \, log_2\left(\frac{1+P_n}{P_n}\right) + log(1+P_n)$$

where tf_x is the frequency of the query term in the top-ranked documents. P_n is given by *F/N* where *F* is the frequency of the query term in the collection and *N* is the number of documents in the collection.

The second useful approximation of the Bose-Einstein statistics is generated by the Stirling formula (Amati, 2003).

Kullback-Liebler divergence computes the divergence between the probability distributions of terms in the whole collection and in the top ranked documents obtained for a first pass retrieval using the original user query. (Cover & Thomas, 1991) For the term t this divergence is :

$$KLD_{(PR,PC)}(t) = \; P_R(t) \, log \frac{P_R(t)}{P_C(t)}$$

where $P_R(t)$ is the probability of the term *t* in the top ranked documents, and $P_C(t)$ is the probability of the term *t* in the whole collection.

QE methods implemented in Terrier results can be seen from the Table 2.

4.2 Proposed QE Results

4.2.1 First Experiment

In the first experiment, we searched each query term in our word-context matrix and calculate cosine similarity ratio with all other words. If this value is less than 0.1 (because we use one minus the cosine in our system) we added this word to the query. We applied the same weights with main query words, and received MAP value of 0.3093.

4.2.2 Second Experiment

As shown in the first experiment, the MAP result is below our baseline. This may be caused by reason that we didn't use any weighting method for the query terms or some expansion terms caused query drifting. In the second experiment, we attempted to develop a hybrid system. We first performed Terrier Query Expansion and got the expanded terms. Then, we used our word-context matrix and we compared these two systems words. If a word is included in both result, we added this word to the query. Also we used the weight taken from QE methods implemented in Terrier for the terms this time. Table 2 summarizes the results that we obtained from experimentation. The second column in each measure, shown in boldface font, indicates the results of our approach for each method. These results show that our approach clearly improves the performance of all three QE methods we tested.

Table 2: *Results of experiments for different methods.*

Method	**MAP**		**R-prec**		**P@10**	
Baseline	0.343		0.361		0.567	
Bo1	0.365	**0.377**	0.380	**0.393**	0.576	**0.589**
Bo2	0.361	**0.371**	0.374	**0.386**	0.565	**0.589**
KL	0.366	**0.377**	0.381	**0.393**	0.574	**0.586**

5. CONLCUSIONS AND FUTURE WORK

In this study, we proposed a Query Expansion (QE) approach on a Turkish Text collection based on word-context matrix with a sliding fixed sized window and Singular Value Decomposition (SVD) method. In order to evaluate the proposed approach, we performed a baseline experiment. Then, we conducted QE methods implemented in Terrier and obtained Mean Average Precision (MAP) results. Finally, we evaluated our proposed Query Expansion approaches.

In the first experiment we expanded the queries with only our system and we couldn't get a good result. But in the second experiment we used a hybrid method (our algorithm plus QE methods implemented in Terrier) and this time we could outperform the baseline and QE methods implemented in Terrier.

We can conclude that co-occurrence information succeeds on synonym and automatically synonym can help us in Query Expansion. Maybe our co-occurrence matrix based technique can be combined to QE mechanism implemented in Terrier in the future.

Currently, we are working on fine-tuning and how to get higher MAP scores from the queries that we couldn't enhance.

For now we work only with single words. But plenty of queries have word phrases. Handling word phrases with the single words, can improve results fairly. To do this, of course a mechanism must be developed to catch word phrases from the query words.

Finally, we don't change term weightings after running Terrier. In other words we use them with the same states. Maybe a work can be carried out to adjust weights. i.e., the importance of words in the query must be determined.

6. ACKNOWLEDGEMENTS

We thank ELIAS for the awarded scholarship for participating in ESSIR 2015 and the FDIA Symposium.

7. REFERENCES

Amati, G. (2003). Probability Models for Information Retrieval based on Divergence from Randomness. PhD thesis, University of Glasgow.

Can, F, Kocberber, S. Balcik, E., Kaynak,C. Ocalan, H.C., Vursavas, O.M. (February 2008), Information retrieval on Turkish texts. Journal Of The American Society For Information Science And Technology, vol. 59, no.3, pp.407-421.

Cover, T.M., Thomas, J.A. (1991) Elements of Information Theory. Wiley-Interscience, New York, USA.

Harris, Z. (1954). Distributional structure. Word, 10(23), 146–162.

Landauer, T.K. Dumais, S.T. (1997). A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review, 104(2), 211–240.

Manning, C. D., Raghavan, P., Schütze, H. (2009) An Introduction to Information Retrieval. Cambridge University Press, Cambridge, England.

S-Space Package. https://code.google.com/p/airhead-research/ (12 April 2015)

The Terrier IR Platform. http://terrier.org/docs/v3.6/ (15 September 2012)

trec_eval. http://trec.nist.gov/trec_eval/ (26 September 2012)

Turney, P. D., Pantel, P. (2010) From Frequency to Meaning: Vector Space Models of Semantics. Journal of Artificial Intelligence Research. 37, 141-188

A Framework for Enhanced Text Classification in Sensitivity and Reputation Management

Graham McDonald
School of Computing Science
University of Glasgow
Glasgow, G12 8QQ
g.mcdonald.1@research.gla.ac.uk

Freedom of Information (FOI) laws state that government documents should be open to the public. However, many government documents contain *sensitive* information that is exempt from release. In this PhD programme, we aim to develop a framework that can automatically classify sensitive information in documents. However, automatic classification of sensitive information is a complex task that requires a relative judgement on the *effect* of a combination of factors. In this paper, we present an overview of the features of sensitivity that we can use to automatically classify documents containing FOI exemptions, such as *International Relations*. Moreover, we argue that current Named Entity Recognition (NER) approaches to classifying sensitive information are not appropriate for classifying FOI exemptions and, therefore, we need classification models that consider the document's *content* and *context* at the *time* of classification.

1. INTRODUCTION

Democratic governments are increasingly following policies of openness and transparency. Moreover, Freedom of Information (FOI)[1][2] laws state that government documents should be open to the public. However, many government documents contain information that is of a *sensitive* nature, such as *personal* or *confidential* information. Therefore, FOI laws make provisions that exempt sensitive information from being released into the public domain. It is essential that all such sensitivities are identified in government documents prior to transfer to the archives. Therefore, the governments of the United Kingdom (UK) and America (USA) have recently recognised that there is a timely need for new algorithms that can detect sensitive information in documents to avoid accidental disclosure (D.A.R.P.A. (2010); Allan (2014)).

In this PhD programme, we aim to develop a framework that can automatically classify sensitive information in documents. However, assessing the sensitivity of information and, moreover, automatically classifying sensitive information is a complex task. For example, in our initial work, we focus on a particular UK FOI exemption, namely *International Relations*, that protects the interests of the UK abroad. This exemption can apply to a document if it contains inappropriate language or content that is potentially reputationally damaging. Therefore, assessing the sensitivity of information requires a relative judgement on the *effect* of a combination of factors.

In the remainder of this paper, we argue that to be able to automatically classify FOI sensitivities, such as International Relations, we need to identify features of sensitive information that relate to three key attributes of sensitivity, namely, the document's *content*, the *context* in which the document was created and the *time* at which we are classifying the document.

2. RELATED WORK

Most research into automatically classifying sensitive information in documents has focused on personal data. Early approaches to document anonymisation came from within the domain of clinical records (Tveit *et al.* (2004)) and used medical dictionaries for term-matching or regular expressions for pattern identification (Sweeney (1996)). However, these approaches were costly, fragile and restricted in their application generalisability. Therefore, recent research into classifying sensitive information has tended to focus on more generalisable approaches.

[1] http://www.legislation.gov.uk/ukpga/2000/36/contents
[2] http://www.foia.gov

© *The Author(s). Published by BCS Learning and Development Ltd.*
Proceedings of the 6th Symposium on Future Directions in Information Access 2015

Named Entity Recognition (NER) is a popular general approach for detecting sensitive information in documents. For example, Abril *et al.* (2011) adapt approaches from Statistical Disclosure Control (Willenborg and De Waal (2001)) and Privacy-Preserving Data Mining (Agrawal and Srikant (2000)) to mask named entities in documents. However, in these approaches all named entities are considered sensitive and, therefore, applying NER masking can reduce a document's utility. With this in mind there has been a shift in the focus of sensitive information classification from simple NER redaction to document sanitisation.

Document sanitisation aims to produce a privacy-preserved version of a document that retains the original document's utility. Sánchez *et al.* (2012) presented a document sanitisation approach that assumed sensitive text is more specific than non-sensitive text. Using the Information Content (IC) of noun phrases as a measure of the phrase's sensitivity they classified phrases with an IC score above an empirically set threshold β as sensitive. This approach focused on identifying personal information. However, they also identified potentially confidential phrases and showed that their approach has the potential for identifying a broader range of sensitivities than NER approaches.

In our previous work (McDonald *et al.* (2014)), we deployed a text classification approach to classify Personal Information and International Relations FOI exemptions. In that work, we extended the text classification with additional features such as the entities in the document, a country risk score and a subjective sentences count. We achieved promising results for a proof-of-concept, however, to fully address the problem of automatic classification of sensitive information we must consider the three key attributes of sensitivity outlined in Section 1.

3. FEATURES OF SENSITIVITY

Sensitive information in documents can arise from three key attributes of sensitivity. Firstly, a document can contain sensitive content, such as inappropriate language. Secondly, often the sensitive nature of a document is a result of the context in which the document was created and, thirdly, sensitivities emergence and decay over time.

Content: A document's content has many potential sources of sensitivity. Firstly, the topic could be sensitive in its own right. However, sensitivity relating to a topic usually arises from what is said about the topic or how it is said. For example, in a report claiming that Croatia are suspected of having violated an international treaty, discussion of the treaty could be sensitive. However, the discussion of the violation is sensitive since the information could be disputed and potentially damage relations with Croatia. Knowledge of the existence of the treaty can help to highlight the potential sensitivity. However, to detect this sensitivity we need to look more closely at the language used and the structure of the text.

A second posible source of sensitivity is the tone of the language used in reference to an entity, for example commenting that a foreign government is “lazy” or “corrupt”. Moreover, culturally inappropriate or politically incorrect references about significant figures can be deemed sensitive.

Thirdly, information that can give a competitor a strategic advantage can also be sensitive, such as reporting that a country is inadequately prepared for a terrorist attack. This sensitivity is more difficult to detect than it might first appear. The reporting of a terrorism incident, or a government's reaction to terrorism, is not by itself sensitive information. It is the appraisal of the government's ability that causes the sensitivity.

Lastly, the source of information is significant in deciding if the information is sensitive. For example, information that has been supplied in confidence is sensitive, however, reporting information from a press conference is not.

To automatically classify the content of sensitivities such as International Relations, we need to identify sensitivity-specific language constructs, such as sequences of terms or parts-of-speech, that are indicative of the sensitivity and use the identified vocabularies to train sensitivity-specific classifiers.

Context: The context in which a document is created is important for sensitivity for two main reasons. Firstly, documents created in a particular context, such as by the *same author* or in a particular *date range*, can discuss related or similar content. Therefore, clusters of related sensitivities can exist. For example, documents produced by a particular government department within a certain date range are likely to produce many documents on a topic. The sensitivities associated to a particular topic are likely to share certain attributes and features. Therefore, we should be able to better identify the sensitivities relating to certain batches of documents if they are viewed within the context that they were created.

Secondly, sensitivities can span multiple documents. Moreover, the sensitive nature of one document might not be apparent without viewing other related documents. To address this we can classify documents from within the same context and propagate (potential) sensitivities to related documents to see if inter-document sensitivities exist. Moreover, by training context-specific classification models we expect to better identify context-dependent sensitivities.

Time: Sensitivities evolve and decay at varying rates. Moreover, the duration of existence for some types of sensitivity are not well defined. For example, documents accounting the sinking of the Argentinian war ship ARA General Belgrano in the Falklands War of 1982 were considered highly sensitive for many years after the event. However, many of details in these documents are now freely available. Therefore, to effectively classify sensitivity in documents over time, the classification models must be able to adapt to the changing vocabulary of currently sensitive content. External resources can help to identify information that is in the public domain. However, as previously outlined, sensitivity arises from the specific aspects of the topics being discussed and this increases the complexity of the task.

4. IMPLEMENTATION

In our initial work (McDonald *et al.* (2015)) we have looked at statistical methods for automatically identifying sensitive content that relates to information supplied in confidence. More specifically, we found that by identifying part-of-speech n-grams that are specific to this sensitivity, we can use the identified n-grams to train sensitivity-specific classifiers. Moreover, we found that this approach can achieve markedly improved recall of this sensitivity compared to a recent approach from the literature, that has been shown to achieve high levels of recall of sensitive text in other domains.

To further develop this work we aim to answer three main research questions: RQ1 What are the most effective methods for automatically identifying sensitivity-specific language constructs? RQ2 How can sensitivity-specific vocabularies be constructed and maintained to be effective for sensitivity classification over time? and RQ3 What is the impact of training context-specific classification models on sensitivity classification?

5. CONCLUSIONS

In this paper, we have presented an overview of sensitivity relating to FOI exemptions such as International Relations. We have argued that to successfully classify these sensitivities we need to go beyond the current NER based approaches. Moreover, we need to develop classification models that can identify features of sensitivity relating to the document's content, the context in which the document was created and the current sensitivities at the time of classification. More specifically, we argue that effective classification of sensitive documents can be achieved by constructing sensitivity-specific vocabularies from language constructs, such as sequences of parts-of-speech, that are specific to individual sensitivities. Moreover, we can use the identified vocabularies to train effective classifiers that can identify passages of sensitive text in documents. Furthermore, by training context-specific classification models we will be able to better identify inter-document sensitivities.

6. ACKNOWLEDGEMENTS

We thank Dr Iadh Ounis and Dr Craig Macdonald for supervising this PhD programme and the ELIAS ESF Research Networking Programme for the scholarship award to participate in ESSIR 2015 and the FDIA Symposium.

REFERENCES

Abril, D and NA, G and Torra, V (2011) *On the Declassification of Confidential Documents* Modeling Decision for Artificial Intelligence. Springer.

Agrawal, R and Srikant, R (2000) *Privacy-preserving data mining*. ACM Sigmod Record. Vol 29.

Allen, A (2014) *Records Review*. UK Government. https://www.gov.uk/government/publications/records-review-by-sir-alex-allan.

Defense Advanced Research Projects Agency (2010) *DARPA, New technologies to support declassification*. Request for Information (RFI).

McDonald, G and Macdonald, C and Ounis, I and Gollins, T (2014) *Towards a Classifier for Digital Sensitivity Review*. In *Proc of ECIR*.

McDonald, G and Macdonald, C and Ounis, I (2015) *Using Part-of-Speech N-grams for Sensitive-text Classification*. In *Proc of ICTIR*.

Sánchez, D and Batet, M and Viejo, A (2012) *Detecting sensitive information from textual documents: an information-theoretic approach*. Modeling Decisions for Artificial Intelligence. 173-184.

Sweeney, L (1996) *Replacing personally-identifying information in medical records, the Scrub system.* In *Proc. of AMIA*.

Tveit, A and Edsberg, O and Rost, TB and Faxvaag, A and Nytro, O and Nordgard, MT and Ranang, MT and Grimsmo, A *Anonymization of general practitioner medical records*. In *Proc of HelsIT*.

Willenborg, L and de Waal, T (2001) *Elements of Statistical Disclosure Control*. LNCS. Vol 155. Springer-Verlag New York.

Different tools for handling Geographic Information Retrieval problems

Yisleidy Linares Zaila
PhD. Student Computer Science and Engineering
University of Bologna
Mura Anteo Zamboni 7, BO 40138
Italy
yisleidy.linares2@unibo.it

Usually in natural language several amounts of geographic references are found, this comes with the common necessity of giving a context with time and location details. Considering these locational semantics, user queries may be satisfied in a more accurate way. In this work a geo-ontology is built for identifying geographic terms. A toponym disambiguation algorithm is proposed for assigning to a place name its corresponding location. The importance of geographic terms in documents is determined by means of a weighting strategy, that is also used to compare geographic contents and to provide a ranking of results. It is also provided a technique for combining a standard textual ranking and the obtained geographic ranking. Final results are evaluated using GeoCLEF test collection and baseline techniques.

geographic information retrieval, toponym disambiguation, spatial similarity measure

1. INTRODUCTION

It is estimated that more than 70% of all information in the world has some kind of geographic features (Jones et al. (2004)). Considering that users queries with geographic references are very natural, the development of search engines aware of geographical semantics has received lots of attention in both the academic and the industrial aspects.

According to Abdelmoty et al. (2005) one definition of Geographic Information Retrieval is the provision of facilities to retrieve and rank by relevance documents or other resources from an unstructured collection, on the basis of queries specifying both theme and geographic scope. This definition carries some main challenges such as:

i. Identification of geographic terms in documents and associating these terms with appropriate geographic locations: A common problem in GIR is that different locations may be named in the same way, this problem is called *toponym disambiguation*. Typical approaches to toponym disambiguation include using knowledge-based , map-based, data-driven methods, or a combination of the three (Buscaldi (2011)). Knowledge-based and data-driven approaches usually incorporate toponym-related information which is used to derive disambiguation rules (e.g. SPIRIT Jones et al. (2004)) and to train machine learning algorithms (Martins and Calado (2010)) respectively. Map-based methods assign geographic locations to toponyms by taking into account the spatial distribution among them (Leidner (2004, 2007)).

ii. Development of spatial similarity measures for comparing geographic information: The similarity between geographic terms is often approximated by geometric or geographic measurements, such as Euclidean distance, overlap or direction (Larson and Frontiera (2004)).

iii. Techniques to properly combine geographic and thematic relevance: A known technique combines documents from different rankings according to their scores, showing that outperforms other approaches in a IR context, as well as in a GIR context (Lee (1997); Palacio et al. (2010)).

In this work we proposed the GeoNW ontology, which is based on GeoNames, WordNet and Wikipedia resources. This stores all geographic knowledge that is used for extracting, analysing and comparing the geographic content present in documents and queries. The toponym disambiguation problem is handled by using the information

© Yisleidy Linares Zaila. Published by BISL.
Learning and Development Ltd.
Proceedings of the 6th Symposium on
Future Directions in Information Access 2015

stored in GeoNW. It is based on the assumption that in a document, the geographic location with more neighbours also present has a higher probability to be the one actually referenced. A weighting strategy for quantifying the degree of influence of a geographic term over a document is also proposed. It is based on the frequency of the geographic term in the document and on its hierarchical and topological relation with the other geographic terms found in the text. Queries are similarly processed and a new spatial matching function to measure the geographic similarity between a document and a query is applied. It uses the results obtained from the weighting strategy. Moreover, we define a strategy for combining standard textual similarity measures with our geographical similarity measure. The proposed algorithm classifies as more relevant documents those with higher rank in both textual and geographic rankings. As a final result a unique ranked list of documents is obtained, expecting it better satisfies the user needs. Finally, the approach is evaluated using GeoCLEF (Mandl et al. (2009)) test collection and results are compared with baseline techniques.

2. GIR TOOLS

A GIR architecture can be seen as a model that separately analyses thematic and geographic information present in texts. The output is a ranked list of documents which are relevant to a specific user query (Figure 1).

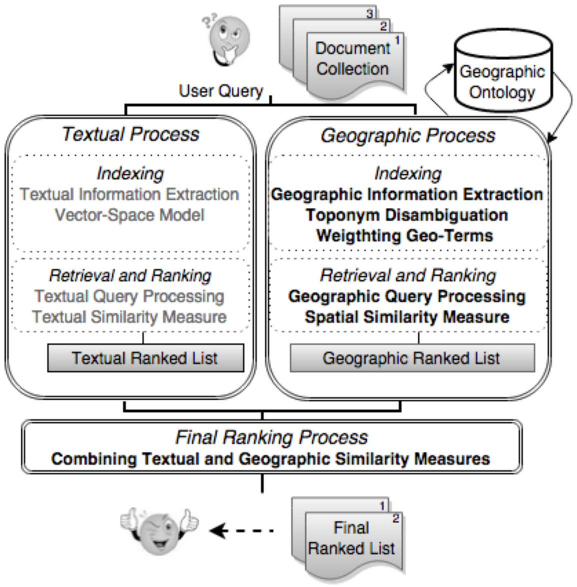


Figure 1: *GIR architecture*

In this work, tasks corresponding to *Textual Process* are achieved using standard IR techniques. The main contributions lie on *Geographic Process*. First, toponyms are extracted and disambiguated using GeoNW ontology. Then, to each geographic term in a document a weighting value that represents its importance in the document is assigned. The geographic information between a query and a document is compared through a spatial similarity measure obtaining a geographic ranking list of documents which is combined with a textual ranking list providing the final result that is showed to the user.

2.1. GeoNW

GeoNW is the geographic ontology, which is built using three different sources: GeoNames, WordNet and Wikipedia. One of the main reasons for building a new geographic knowledge source was that existing databases such as GeoNames refer to several places with the same geographic location (i.e. *Boston*) increasing the ambiguity problem. It means, there are many entries corresponding to the same place.

The structure of GeoNW is shown in Figure 2. Each physical or administrative place has associated a set of synonyms that were obtained from the alternateNames in GeoNames and from synsets in WordNet. This relation allows to recognise *capital of Spain* as *Madrid* or *the city of Light* as *Paris*. Also, each place has its geographic coordinates, which were provided from GeoNames. Furthermore, using Wikipedia a set of nationalities adjectives were added to administrative places, allowing to relate *Italian rivers* with *rivers in Italy*.

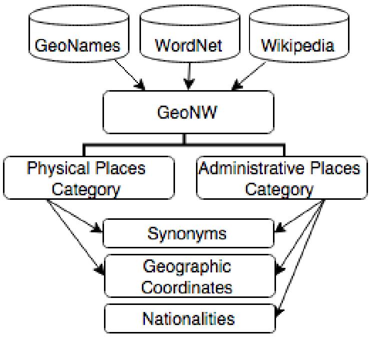


Figure 2: *GeoNW*

2.2. Toponym Disambiguation

According to Leidner (2007), toponym ambiguity can be classified in: i) morpho-syntactic ambiguity (e.g. *nice* the English adjective and the city in France); ii) feature type ambiguity (e.g. *Mississippi* the State of USA and *Mississippi* the river); iii) referential ambiguity (e.g. *London* the city in England, UK or the city in Ontario, Canada). Our toponym disambiguation technique is based on this idea of ambiguity classification.

The goal of the disambiguation algorithm is to assign to a place name a unique geographic element. Thus, each step attempts to remove elements from the candidate list in order to take the one more related to the content. The technique uses the information stored in GeoNW and assumes that the more related a geographic entity is to other geographic terms in the document, the more probable to be referenced it will be.

If the place name can be a physical place (i.e. river, mountain) or an administrative place (i.e. country, city, town), a Feature Type Disambiguation strategy is applied. It is based on the candidate feature type[1] frequency in the document. The technique takes the geographic reference that corresponds to the more mentioned feature type.

The case when we have the same toponym associated to different geographic entities is solved through the Referential Disambiguation technique. In this case, ambiguity is only related to administrative places. The algorithm uses the hierarchical relationships among the ambiguous place name and the other geographic locations present in the document. This hierarchical relationship is obtained from GeoNW. The method chooses the place which its geographic location is closer, in the hierarchy, to all other places present in the document.

2.3. Document Geographic Focus Detection

The geographic focus is represented as a set of pairs $< w_i, g_i >$ where g_i is a geographic term and w_i is a numeric value that represents its influence over the document. It uses the relationships among all georeference in the document. It is also based on the principle that more relatives a geographic term has in a document, more the geographic term is associated to the document.

The weighting value assigned to each geographic term in a document d_i is computed by the expression:

$$\frac{freq(gt)}{maxG} * \left(TI(gt, d_i) + DI(gt, d_i) + \frac{1}{|dG_i|} \right)$$

where TI (Topologic Influence) and DI (Distance Influence) are functions that evaluate topological and metrical relationships respectively, $freq(gt)$ is the frequency of gt in d_i, $maxG$ is the maximum value reached by the expression above and $|dG_i|$ is the number of geographic elements in d_i. TI function is mainly based on the number of common ancestors between two geographic references, while DI is based on the geographical distance between two locations that have at least one common ancestor. If two geographic terms do not have any common ancestor, DI output is infinity. The latter avoids the negative effect of not related geographic terms that are too far from gt.

2.4. Query Geographic Focus Detection

Due to, queries are usually composed by a short number of words, and usually have only one geographic term, the disambiguation process can not be the same used for documents. Given $cList$, defined as the list of all possible alternatives of an ambiguous place name tp. The list $cList_w$ is defined as:

$$cList_w = \left[< w_1, c_1 >, ..., < w_n, c_n > \right]$$

where $c_i \in cList$ and w_i is the ratio of the number of times that tp matches with c_i and the total frequency of c_i in the collection D. Notice that $cList_w$ is built during the geographic indexing and it depends on the toponym disambiguation process explained above.

2.5. Spatial Similarity Measure

The proposed technique for assigning a score to a document according to a query is very intuitive. It is directly based on the results of processing the geographic focus of the query Q and documents $d_i \in D$. It can be seen as a combination of distance and topological methods, because it is strongly related to the relationships among the geographic terms and the geographical distance among them. Let QG and dG_i be the sets that represent the geographic focus of the query Q and the document d_i respectively, the spatial score function S_G is defined as:

$$S_G(QG, dG_i) = \sum_{n=1}^{l} \sum_{m=1}^{k} (w_n * w_m)_{c_n = g_m}$$

where c_n is a geographic entity present Q and g_m is a geographic entity present in d_i. Computing this expression for all documents in D, the geographic ranked list that corresponds to Q is obtained.

2.6. Combining Textual and Spatial Similarity Measure

Let R_T and R_G be the textual ranked list and geographic ranked list of the query Q in the collection D. R_T and R_G contains normalised values in a range of $[0, 1]$, where 1 is the score of the most relevant document in D to the query Q. The CombTG function is defined as:

$$CombTG(R_T, R_G) = \sum_{i=1}^{k} \beta * (st_i + sg_i)$$

[1] Feature types correspond to physical places, such as rivers, mountains, lakes, etc.

where st_{π_i} and sg_{ϕ_i} are the textual and geographic score of document d_i and β takes value 2 if d_i is in R_T and R_G, 1 if d_i is only in R_T and 0.5 if d_i is only in R_G. This strategy benefits documents retrieved in both lists and penalizes those that were retrieved only by their geographic information. It is based on the assumption that documents whose textual information is not relevant to the query will have less importance than those that do are.

3. EVALUATION

The algorithms are implemented as an extension of Terrier[2] tool. The evaluation process is based on GeoCLEF[3] test collection. There are 166 477 documents and 25 geographic topics. GeoCLEF documents were extracted from a set of articles from *The Los Angeles Times (1994)* and *The Glasgow Herald (1995)*.

A first evaluation of our approach was made by comparing with Terrier baseline approaches. The best results of Terrier baseline algorithms were obtained by using $DLH13$ (Lu et al. (2013)) and LGD (Clinchant and Gaussier (2011)) models. Also we use the $BM25$ approach as it is a well known standard technique for retrieval applications. In Table 1, the first three rows correspond to the results obtained by the application of standard IR techniques without geographic analysis, while the last three rows show the results using these techniques for textual analysis and the proposed strategy for geographic analysis. As a promising result we can see that our approach improves all MAP values using for textual processing any of the weighting models mentioned above.

Model	Recall	MAP
$BM25$	0.905	0.378
$DLH13$	0.908	**0.404**
LGD	**0.919**	0.394
$CombTG\_BM25$	0.902	0.465
$CombTG\_DLH13$	0.902	0.487
$CombTG\_LGD$	**0.909**	**0.489**

Table 1: *Overall results*

On the other hand, results reported by GeoCLEF 2008 track overview Mandl et al. (2009) using MAP evaluation measure are around 0.3, thus our strategy reaches better results.

3.1. Toponym Disambiguation Strategy Evaluation

Toponym disambiguation is one of the problems we attempt to solve in this work. For evaluating the proposed technique we compare with a naive strategy that identifies as the geographic location of an ambiguous place name the one with the largest population. In Figure 3 $CombTG\_BM25\_pop$, $CombTG\_DLH13\_pop$ and $CombTG\_LGD\_pop$ correspond to the behaviour of our approach using for textual analysis $BM25$, $DLH13$ and LGD weighting models respectively and for disambiguating toponyms the geographic location with the largest population; while $CombTG\_BM25$, $CombTG\_DLH13$ and $CombTG\_LGD$ correspond to the same strategies but using the proposed disambiguation technique. As it is shown for all cases using our disambiguation algorithm, a slightly better result is obtained.

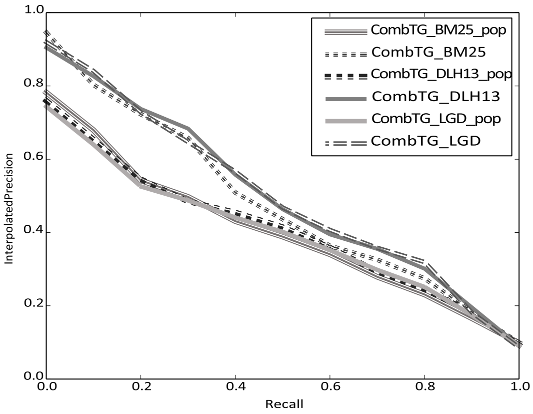


Figure 3: *Evaluating Toponym Disambiguation Strategy*

Although a more precise evaluation of the toponym disambiguation technique is needed in order to compare with other toponym disambiguation strategies, these preliminary results show that our approach produces a positive effect on the final result.

4. CONCLUSIONS

This paper describes a new approach for retrieving and ranking documents according to textual and geographic information. It uses standard IR techniques for processing textual information, while the geographic information is analysed using the geographic ontology GeoNW. A toponym disambiguation method is proposed, which according to the experiments, improves the naive technique. This paper also proposes a weighting strategy in order to quantify the influence of a geographic entity over a document. It is based on topology and distance relationships among geographic terms. Using GeoCLEF

[2] http://terrier.org/
[3] http://www.uni-hildesheim.de/geoclef/

test collection, the strategy is evaluated and results outperform baseline techniques. The best result is achieved by combining the LGD model with the proposed geographical analysis, obtaining a mean average precision of 0.489. Currently we are working on a more exhaustive evaluation of our strategy comparing the results with other approaches. We are also planning to include query expansion techniques in order to continue improving the overall results.

ACKNOWLEDGEMENTS

I thank my supervisor Danilo Montesi for his support in the development of this work. Special thanks also to ELIAS (Evaluating Information Access Systems) ESF Research Networking Programme for the awarded scholarship to participate in the 10th European Summer School in Information Retrieval (ESSIR 2015) and the FDIA Symposium.

REFERENCES

Abdelmoty, A. I., P. D. Smart, C. B. Jones, G. Fu, and D. Finch (2005). A critical evaluation of ontology languages for geographic information retrieval on the internet. In *Journal of Visual Languages and Computing*, Volume 16, pp. 331–358. Elsevier.

Buscaldi, D. (2011). Approaches to disambiguating toponyms. *SIGSPATIAL Special 3*(2), 16–19.

Clinchant, S. and É. Gaussier (2011). Retrieval constraints and word frequency distributions a log-logistic model for ir. *Information retrieval 14*(1), 5–25.

Jones, C. B., A. I. Abdelmoty, D. Finch, G. Fu, and S. Vaid (2004). The spirit spatial search engine: Architecture, ontologies and spatial indexing. In *Geographic Information Science, Third International Conference, GIScience, Adelphi, MD, USA*, Volume 3234 of *Lecture Notes in Computer Science*, pp. 125–139. Springer.

Larson, R. R. and P. L. Frontiera (2004). Spatial ranking methods for geographic information retrieval (gir) in digital libraries. In *Proceedings of the 8th European Conference on Research and Advanced Technology for Digital Libraries (ECDL)*, Volume 3232 of *Lecture Notes in Computer Science*, pp. 45–56. Springer-Verlag.

Lee, J. H. (1997). Analyses of multiple evidence combination. In *ACM SIGIR Forum*, Volume 31, pp. 267–276. ACM.

Leidner, J. L. (2004). Towards a reference corpus for automatic toponym resolution evaluation. In *Workshop on Geographic Information Retrieval, Sheffield, UK*.

Leidner, J. L. (2007). *Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names*. Ph. D. thesis, Institute of Communicating and Collaborative Systems, School of Informatics, University of Edinburgh.

Lu, S., B. He, and J. Xu (2013). Hyper-geometric model for information retrieval revisited. In *Information Retrieval Technology*, Volume 8281 of *Lecture Notes in Computer Science*, pp. 62–73. Springer Berlin Heidelberg.

Mandl, T., P. Carvalho, G. Di Nunzio, F. Gey, R. Larson, D. Santos, and C. Womser-Hacker (2009). Geoclef 2008: The clef 2008 cross-language geographic information retrieval track overview. In *Evaluating Systems for Multilingual and Multimodal Information Access*, Volume 5706 of *Lecture Notes in Computer Science*, pp. 808–821. Springer Berlin Heidelberg.

Martins, B. and P. Calado (2010). Learning to rank for geographic information retrieval. In *Proceedings of the 6th Workshop on Geographic Information Retrieval*, pp. 21. ACM.

Palacio, D., G. Cabanac, C. Sallaberry, and G. Hubert (2010). On the evaluation of geographic information retrieval systems. *International Journal on Digital Libraries 11*(2), 91–109.

Query Recommendation as Query Generation

Matthew Mitsui
Department of Computer Science
Rutgers University
110 Frelinghuysen Road
Piscataway Township, NJ 08854, USA
mmitsui@cs.rutgers.edu

Researchers have studied query recommendation to address various aspects of the user search experience. Several contributions in this area use search logs to recommend existing queries from the log, using query-based similarity metrics or log-based probabilities. However log-based recommendations are limited to queries issued by users, not necessarily utilising the full potential of a search system. The proposed work here intends to approach query recommendation from a generative perspective. It proposes to generate novel queries by using search logs and web crawls to model a user's knowledge and to recommend queries to satisfy knowledge deficiencies.

query recommendation, information retrieval, IR, exploratory search, recommender systems, web crawl

1. INTRODUCTION

The term "query recommendation" refers to several types of assistance tools that aid users during a search session. In general these tools are back-end or front-end processes that issue or recommend search engine queries during a user's search session. Back-end recommendation algorithms require little user involvement and re-rank search results. Wang et al. (2015), as an example, incorporates the most common follow-up query results to a current query, to re-rank results and save user effort. Front-end recommenders require more user involvement but are potentially more informative for users. Spelling correctors are informative in an obvious way, and "related searches" offered by Google, Yahoo! and Bing attempt to inform users how to construct more effective queries (Sisode and Patil (2014)).

Most current query recommendation techniques are designed for personalisation and rely predominantly on **search logs** as input data. A search log is a set of users' search sessions, with each session containing information about a user's sequence of queries, clicks on search engine result pages (SERPs) and possibly time spent on results. This data can provide per-query implicit relevance feedback that is useful for measuring query satisfaction, page usefulness and pairwise query similarity scores - key components for recommendation algorithms. While search logs have been used to make simple, powerful recommenders the resulting algorithms are limited to providing previously issued queries and previously discovered pages from the logs. What if we wanted to recommend rarely-discovered, topic-relevant information in a front-end recommender? This is somewhat possible by modeling long-tail queries from a log, but users do not always know how to issue queries well (Taneja and Chaudhary (2012)). It eventually becomes necessary to generate novel queries to fulfill information deficiencies and even teach users how to best construct a search engine query. For this it becomes necessary to model words, pages, topics and their relationships, requiring a larger scale corpus such as the Web.

Can a recommender construct novel queries to effectively and efficiently retrieve rare, useful information? Furthermore how would such queries compare to those currently issued by users? This paper explores the opportunities and challenges presented by recommending novel queries by mining search logs and a web crawl. We situate our work within the literature of query recommendation and web document modeling, and we give possible directions for technical contributions.

2. BACKGROUND

2.1. Query Recommendation

Query recommendation has a long history dating to at least the early 2000s, yielding various techniques

© The Author. Published by BCS
Learning and Development Ltd.
Proceedings of the 6th Symposium on
Future Directions in Information Access 2015

with differing applications. Most approaches harness some aspect of search logs to make personalised recommendations to users. These techniques can utilise sequences of queries in a log, click-through information on SERP URLs, the snippets in SERP results and the full text of linked pages. Baeza-Yates et al. (2004), for instance, use click-through information to cluster queries by similarity, offering similar substitutes when a current query fails to yield good results. An opposing application from Vahabi et al. (2013) similarly uses query distance to recommend orthogonal queries - dissimilar queries that are similar enough to be topically relevant. He et al. (2009) and Boldi et al. (2008) model n-grams and even general query graphs respectively, lending themselves to the applications in the Introduction.

While our approach will similarly use search logs - as they are indispensable for personalisation - it will not directly derive the surface query (e.g. “new movies”) from the logs. Our approach will use logs to determine what a user has already seen but will automatically generate queries to satisfy some deficiency in the user’s knowledge. This deficiency will be determined with a web crawl. While we will recommend orthogonal queries like Vahabi et al. (2013) their queries are directly derived from logs.

2.2. Modeling the Web

As we will use web crawls to model users’ knowledge and to build queries we must overview representations of web documents. All representations have their basis in classical information retrieval (IR) systems. These IR-based representations consist of documents that have been preprocessed through several steps, such as removal of formatting tags (e.g. HTML user tags) and stopwords (i.e. common non-content words like “the”). Documents are also stemmed or lemmatised - that is words are converted to a basic root form. After preprocessing they are treated as bags of words and converted into weighted vectors of terms. Researchers have applied various weighting schemes to document text, which are functions of the frequencies of terms in a document and/or their frequencies in a corpus. One scheme, for instance, is term frequency inverse document frequency (TF-IDF), in which words’ weights are a product of their frequency in a document and the inverse of their frequency in a corpus. Some Web-IR methods extend the IR-based methods by additionally weighting terms according to their HTML tags - e.g. whether they are in the page’s body or title. Dimensionality reduction techniques like Latent Semantic Indexing are sometimes applied to documents as well (Micarelli et al. (2007)), in order to reduce the number of dimensions and to extract the most meaningful ones.

The former representations yield matrices of documents and their respective term weights. Other representations add structure to the web. Artificial Neural Networks use layers of nodes to represent query terms, document terms and documents, strictly for ranking documents for queries. Bayesian networks similarly connect terms, documents and even topics into a directed acyclic graph for result ranking. Semantic networks represent the relationships between linguistic concepts and can be used to convert a document from a bag of words into a bag of concepts (Micarelli et al. (2007)). WordNet, for instance, groups words into sets of synonyms and can be used to reduce sparsity in simple bag-of-words vectors. Topic models like Latent Dirichlet Allocation reduce dimensions and graphically link words and topics, modeling words and documents as mixtures of topics (Blei et al. (2003)).

One purpose of these models is simply for ranking: to convert documents and queries into a TF-IDF representation and rank them according to their cosine similarity score against an input query, perhaps after preprocessing. Another is to reduce duplication, whether to reduce duplication of documents (by clustering) or to simplify a corpus into relationships between words, topics, themes. Since we will generate queries from the ground up from web pages, we are interested in the latter purpose. We will link documents or passages thematically and count the frequency of terms/concepts to generate queries. We will begin with the most basic approach and incrementally add layers of complexity.

2.3. Exploratory Search

Research in search ranking and recommendation has covered a large variety of search tasks. These tasks include simple fact-finding tasks that can be satisfied in one query, such as navigation to a known web page, retrieving a dictionary definition or finding important facts and relationships involving a historical figure. Our approach is most suited to **exploratory search tasks**. These tasks are so complex as to require a long span of queries and potentially multiple search sessions. They include vacation planning, report writing or synthesising nutritional and exercise information for a holistic wellness program. Such tasks have been estimated to comprise 10% of search sessions and 25% of overall queries (Donato et al. (2010); Kotov et al. (2011); Rose and Levinson (2004)).

Exploratory search tasks are also associated with cognitive behaviors like learning and sense-making. Exploratory searchers develop mental models of their search topic through-out the process of searching. Exploratory search tasks are also open-ended and multi-faceted. Sometimes a user’s information need is ill-structured or is composed of multiple subtopics (Wildemuth and Freund (2012)). As such there is much opportunity to offer users a diverse set of queries that are all possibly useful to their task, particularly queries that take advantage of the user's deficiencies in knowledge in a session spanning multiple queries. While there is considerable interest in exploratory search tasks our work fills a niche that has not been filled by past exploratory search research and search recommendations in this area.

3. METHODOLOGY

Our model for recommendation will operate using two main sources of data: a web page crawl and a search log. We will use search logs from exploratory search sessions of real users, containing queries, their results and the content pages of the query result URLs. We will also include timestamps of clicks and page views as well as implicit relevance feedback - like viewing time - where such data is present. While we currently have data sets with all such data we are open to using other existing exploratory search data sets as well. Training and testing will assume that users are strictly mono-tasking. Search logs will be used to determine a user's topic of interest and to extract relevance feedback on previously viewed information.

We will model the web at various levels of size and complexity. Our simplest approach will be to convert documents into word vectors and to only use documents linked through the task-specific SERPs. When applicable we will grow our set of documents to include those from other comparable exploratory search sessions (e.g. same type of task but different topic) to ensure our method can account for web pages from irrelevant domains without making bad recommendations. We will also increase the complexity, beginning with bag-of-words documents and proceeding to more complex representations like a bags-of-concepts (in a term-based sense as with LSI or as outlined in Baziz et al. (2005)) or other graphical representations of terms and topics. While this will increase the richness, compactness and density of our representations, it will increase the computational overhead of processing web documents. The purpose is to mine a user's relationship to unseen information scattered on the web and to extract salient key words, phrases or passages for recommendation in reasonable time.

Our algorithm will generate a query at various points of the session. We will simulate sessions as if our recommendations are taken by the user at each point of recommendation, until a pre-specified termination condition is reached (e.g. until a specified number of pages and queries have been viewed). We will test single recommended queries and sequences, to test our method's effectiveness as a whole-session tool and a single-query tool.

4. EVALUATION

We will first evaluate the surface-level queries (i.e. the input strings) generated by our algorithm. We will compare intrinsic properties of input queries in actual user logs against those generated by our algorithm - e.g. by comparing their language models (LMs). Comparing LMs can tell us how similar algorithmically generated queries are with those created by real users. We will use other intrinsic measures in this way as well, such as pointwise mutual information (PMI), perplexity and conditional probability of query terms. Similar analyses can be performed on the "expanded form" of the SERP associated with a query, which includes the URLs, search result snippets and the content pages linked by URLs (Metzler et al. (2007)). Several features used by Ashok et al. (2013) to predict success of novels include LMs, part-of-speech tag distributions and distributions of grammar rules and sentiment. While we would not model these for prediction and are comparing arguably smaller documents we could still compare the two sets of queries.

Since our existing data sets have marks of implicit relevance feedback we will initially use existing extrinsically-based measures to quantify the effectiveness of our algorithm's generated queries. Existing measures include normalised discounted cumulative gain (NDCG) and mean reciprocal rank (MRR), which determine whether the queries return useful URLs. A measure that is popular with exploratory search tasks is coverage of relevant web pages. As we expand our web crawl to the entire web we will then navigate away from URL-based measures and adapt them in a content-based way, since our queries are likely to generate previously unseen URLs. We will compare performances at each level of granularity, for instance measuring the coverage of words and concepts. Our analyses will help us determine how well our algorithm captures URLs, words and concepts. We may also need to adapt measures to normalise for multiple suggested queries; some information needs/deficiencies may be so diverse that they cannot be encapsulated within a single query.

We can also include analyses of algorithmic complexity and human assessment. An analysis of the complexity of our model at various stages is necessary to argue for its real-time usability. For human assessment we can give human assessors real queries and simulated queries - or queries in the context of a sequence - to assess them along various rubric criteria (e.g. whether they are on-topic or comprehensible). These assessments can be given via crowdsourcing platforms such as Amazon Mechanical Turk[1], with appropriate pilot testing to test the platform's effectiveness. As our LM-based evaluation metrics cannot determine whether our recommendations are good or bad (just how different they are) human assessment can fill this gap.

5. ACKNOWLEDGMENTS

Special thanks goes to my supervisor, Dr. Chirag Shah, and the paper reviewers for their useful feedback. The work described here was supported by a travel scholarship of the ELIAS (Evaluating Information Access Systems) ESF Research Networking Programme. This work was also supported by the US Institute of Museum and Library Services (IMLS) Career Development grant # RE-04-12-0105-12.

REFERENCES

Ashok, V.G., Feng, S. and Choi, Y. (2013) Success with style: using writing style to predict the success of novels. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (EMNLP'13). 1753-1764.

Baeza-Yates, R., Hurtado, C. and Mendoza, M. (2004) Query recommendation using query logs in search engines. In *Proceedings of the 2004 international conference on Current Trends in Database Technology* (EDBT'04), Wolfgang Lindner, Marco Mesiti, Can Trker, Yannis Tzitzikas, and Athena I. Vakali (Eds.). Springer-Verlag, Berlin, Heidelberg, 588-596.

Baziz, M., Boughanem, M., and Traboulsi, S. (2005) A concept-based approach for indexing documents in IR. In *INFORSID*, (Vol. 2005, pp. 489-504).

Blei, D., Ng, A. and Jordan, M. (2003) Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993-1022.

Boldi, P., Bonchi, F., Castillo, C., Donato, D., Gionis, A. and Vigna, S. (2008) The query-flow graph: model and applications. In *Proceedings of the 17th ACM conf. on Information and knowledge mining* (CIKM'08), New York, NY, USA, ACM, pp. 609618.

Donato, D., Bonchi, F., Chi, T. and Maarek, Y. (2010) Do you want to take notes?: Identifying research missions in Yahoo! search pad. In *Proceedings of the 19th International Conference on World Wide Web* (WWW'10), pages 321-330, New York, NY, USA, 2010. ACM.

He, Q., Jiang, D., Liao, Z., Hoi, S., Chang, K., Lim, E. and Li, H. (2009) Web query recommendation via sequential query prediction. In *Proceedings of the 2009 IEEE International Conference on Data Engineering* (ICDE '09). IEEE Computer Society, Washington, DC, USA, 1443-1454.

Kotov, A., Bennett P., White, R., Dumais, S. and Teevan, J. (2011) Modeling and analysis of cross-session search tasks. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR '11), pages 5-14, New York, NY, USA. ACM.

Metzler, D., Dumais, S., and Meek, C. (2007) Similarity measures for short segments of text. In *Proceedings of the 29th European Conference on IR Research* (ECIR'07). Springer-Verlag, Berlin, Heidelberg 16-27.

Micarelli, A., Sciarrone, F. and Marinilli, M. (2007) Web Document Modeling. In *The Adaptive Web*, Peter Brusilovsky, Alfred Kobsa, and Wolfgang Nejdl (Eds.). Lecture Notes In Computer Science, Vol. 4321. Springer-Verlag, Berlin, Heidelberg 155-192.

Rose, D. and Levinson, D. (2004) Understanding user goals in web search. In *Proceedings of the 13th International Conference on World Wide Web* (WWW '04), pages 13-19, New York, NY, USA. ACM.

Sisode, M.R. and Patil, U.M. (2014) A survey on query recommendation techniques and evaluation of snippet based query recommendation. In *IJCA Proceedings on National Conference on Emerging Trends in Information Technology* NCETIT(1):1-5, December 2014.

Taneja, N. and Chaudhary, R. (2012) Query recommendation for optimizing the search engine results. In *International Journal of Computer Applications.* Volume 50 No.13.

Vahabi, H., Ackerman, M., Loker, D., Baeza-Yates, R. and Lopez-Ortiz, A. (2013) Orthogonal query recommendation. In *Proceedings of the 7th ACM conference on Recommender systems* (RecSys '13). ACM, New York, NY, USA, 33-40.

Wang, J.G., Huang, J., Guo, J. and Lan, Y. (2015) Query ranking model for search engine query recommendation.

Wildemuth, B. and Freund, L. (2012) Assigning search tasks designed to elicit exploratory search behaviors. In *Proceedings of the Symposium on Human-Computer Interaction and Information Retrieval* (HCIR '12). ACM, New York, NY, USA, Article 4, 10 pages.

[1] https://www.mturk.com/

Whether a CQA User is a Medical Professional? Work in Progress

Alexander Beloborodov
Ural Federal University
620002, 19 Mira Street, Ekaterinburg, Russia
xander-beloborodov@ya.ru

This work-in-progress aims to address the problem of detecting whether a CQA user is a medical professional or not. Proposed approach is based on the technique of learning from positive and unlabeled examples. A few classification features and a simple evaluation methodology are presented.

User Expertise, Medical IR, One-class SVMs, Community Question Answering, CQA

1. INTRODUCTION

The paper describes an ongoing research project investigating Community Question Answering (CQA) users answering questions in the domain of people's health. CQA is a service where people are able to ask any question and answer questions asked earlier. There is a wide variety of question topics. Our research project is focused on the people's health topic. It is very important for user, asking question about his/her health, to know the level of expertise of an answering user. We investigate a problem of detecting whether an answering user is a medical professional or not.

The paper is organized as follows: section 2 discusses relevant previous work; the data and resources on which experiments conducted are described in section 3; the prospective method and some features are described in section 4. Finally, in section 5 a few challenges for future work are discussed.

2. RELATED WORK

The tasks of expert search and classification are investigated in a number of works.

In [Liu et al., 2005] the authors focus on automatically finding experts in an open-domain community-based QA service. They cast the expert finding problem as an IR problem where the given question can be viewed as query and the expert profiles can be viewed as documents. Such documents are ranked using language models. While considering primarily language aspects this approach are not taking in account other useful and interesting types of features (for example, statistical or semantic features).

A few different types of features are used in [Pennacchiotti & Popescu, 2011]. The work addresses the task of Twitter user classification by leveraging observable information such as the user behaviour and the linguistic content of the user's Twitter feed. The authors provide an in-depth analysis of the relative value of feature classes and show experimentally that content features are in general highly valuable in classification tasks.

The task of learning from positive and unlabelled examples is considered in [Manevitz & Yousef, 2002]. In [Zhang & Zuo, 2008] a good implementation called One-Class SVMs is suggested and described in detail. This implementation is chosen for our experiments. In addition, the authors argue that the absence of negative information entails a price, and one should not expect as good results as when they are available. That is why it is so important to find a way to form a sample of negative examples.

3. DATA & RESOURCES

The research project is focused on Russian language Community Question Answering (CQA) Otvety@Mail.Ru (Otvety means answers in Russian). All questions in the service are divided into predefined set of categories. We study all questions related to people health with corresponding answers from 4 categories: *Diseases and Medicines; Doctors, Clinics, and Insurance; Doctors' answers; and Kids' Health* in the timespan from 1 April 2011 to 31 December 2012 – 227,828 questions at all.

© Alexander Beloborodov. Published by BCS Learning and Development Ltd.
Proceedings of the 6th Symposium on Future Directions in Information Access 2015

For evaluation purposes the online survey among most active users is conducted. The respondents answered a few questions about their professional skills and motivation to answer questions. Every respondent was able to specify his/her email by which he/she could be identified in the service. Among other questions, people were asked also a question:

Is your profession related to the people's health?

An invitation to participate was sent to about 700 users which are most active in the categories. 171 users participated in the survey, 54 of them specified their emails, 26 respondents answered that their profession is related to people's health, and 28 respondents answered that their profession is not related to people's health.

4. PROPOSED APPROACH

4.1 Training set

Proposing approach is an automatic classification using supervised machine learning that needs sufficiently large training set. Unfortunately, online survey results do not provide us with data set of suitable size. Therefore training set is collected by other means while the online survey results are supposed to be a test set.

CQA answerer has an option to specify his/her answer source. The answer source field has an unstructured plain text format. Users in the health-related categories are often filling the answer source field with phrases such as "I'm physician myself!" or "My medical degree". From 1867 unique answer sources 182 were selected as items certifying that answer author has profession related to people's health. Assuming that source fields contain true statements we have 263 users related to people's health to form positive samples for the training set.

4.2 Proposed method

Although the answer source field often contains medical profession mentions, there are no mentions of other professions the field, so we cannot form a negative samples subset in such a way. Therefore, at this stage, the technique of learning from positive and unlabelled examples is used. One-class SVM with non-linear kernel RBF is used as the technique implementation.

Initial experiments show that the algorithm is often classifying users who specified his/her profession as not related to people's health as positive cases (26 from 28).

4.3 Features

Three types of features are selected for the classification task.

4.3.1. Statistical features

This group of features usually characterizes texts in general. Some examples are mean answer length (in words), an amount of unique words in all user answers, an amount of unique words divided by a whole amount of words answerer mentioned. Additionally we slightly exploit CQA structure to enrich feature set with an amount of user answers divided by an amount of his/her questions inside one particular category.

4.3.2. Linguistic features

After manual investigation of surveyed user answers we assume that speech of a person who is a prospective physician is less emotional than speech of a CQA user in general. According to this assumption we exploit following features: the amount of all general punctuation marks, the amount of "emotional" marks ('!', '?', '(', ')'), the amount of repeating marks ("…..", "!?!?!?!?", ")))))"), fully uppercased words ("EXAMPLE") abuse.

4.3.3. Semantic features

The most interesting features are semantic. Assuming the fact, that a vocabulary of user with medical degree is full of specific terms, a medical domain term dictionary is collected. Feature is an amount of specific terms in user answers.

Many physicians even answering in the CQA do not believe that answer could replace regular real doctor visits. Therefore a hypothesis that users with medical degree more often route a questioner to a real physician is formulated. To verify it two sets of words collected: words analogous to word "physician" and modal words. Examples of "physician"-like words: *doctor, specialist, ambulance, clinic, dentist,* etc. Examples of modal words are *consult, go, see, necessarily, need,* etc. An event when "physician"-like word and modal word appeared in the distance 2 or less from each other are recognizing as recommendation to visit a doctor:

Q: *Help me please. Inexplicable rash on the body! [Photo attached]*

A: *It looks like herpes. It is contagious and can be transmitted from animals. Urgently need to* ***[see a doctor]***

Traditional medicine and ICD-10 drug mentions are serving as features as well.

5. CHALLENGES

There are some challenging questions that we faced during the research project.

The current evaluation methodology is supposed to test the method performance using the online survey results as a gold standard. This approach has a number of drawbacks:

a) We have only 26 respondents answered that their profession is related to people's health, and 28 respondents answered that their profession is not related to people's health. Such amount seems too small to make reliable conclusions but it still useful on the initial stage of the research.

b) A manual investigation shows that answerers, who specified his/her profession as not related to people's health, are often giving answers of a good quality. This can make it difficult to understand who can be considered as a medical professional.

Other important features that need to be integrated into the evaluation process are readability and understandability of expert answers. As shown in [Zuccon & Koopman, 2014], the understandability is considering as a critical issue for supporting online consumer health search because consumers may not benefit from health information that is not provided in an understandable way; and the provision of unreliable medical condition or treatment may led to negative health outcomes.

As follows from [Zhang & Zuo, 2008], with the absence of negative information we should not expect as good results as when negative examples are available. At this point building such a training set is a challenge as well.

One more method drawback is a lack of user-level features, such as the user score, age, gender, and so on.

6. ACKNOWLEDGMENTS

This work was supported in part by the ELIAS (Evaluating Information Access Systems), an ESF Research Networking Programme, and in part by the Russian Foundation for Basic Research, project #14-07-00589 “Data Analysis and User Modelling in Narrow-Domain Social Media”.

The author also would like to thank the anonymous reviewers for their valuable comments and suggestions to improve the quality of the work.

7. REFERENCES

Liu, X., Croft, W. B., & Koll, M. (2005). Finding experts in community-based question-answering services. Proceedings of the 14th ACM international conference on Information and knowledge management (pp. 315-316). ACM.

Pennacchiotti, M., & Popescu, A. M. (2011). A Machine Learning Approach to Twitter User Classification. ICWSM, 11, 281-288.

Zhang, B., & Zuo, W. (2008). Learning from positive and unlabelled examples: A survey. Information Processing (ISIP), 2008 International Symposiums, 650-654. IEEE.

Manevitz, L. M., & Yousef, M. (2002). One-class SVMs for document classification. the Journal of machine Learning research, 2, 139-154.

Zuccon, G. & Koopman, B. (2014). Integrating understandability in the evaluation of consumer health search engines. Medical Information Retrieval (MedIR) Workshop, 11 July 2014, Gold Coast, Australia.

Explanatory opinions: to whom or what is all the fuzz about?

Filipa Peleja
Yahoo Labs
Barcelona
Spain
peleja@yahoo-inc.com

Ioannis Arapakis
Yahoo Labs
Barcelona
Spain
arapakis@yahoo-inc.com

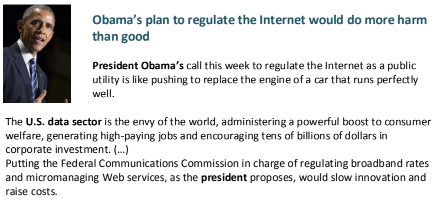
João Magalhães
NovaLincs, Dep. de Informática
Universidade Nova de Lisboa
Caparica, Portugal
jm.magalhaes@fct.unl.pt

Exploiting sentiment relations to improve the accuracy of sentiment analysis has caught the interest of recent research. When expressing their opinions, users apply different sentence syntactic constructions styles. This analysis leverages on a sentiment lexicon that includes general sentiment words that characterize the overall sentiment towards the targeted named-entity. However, in most cases, target entities are themselves part of the sentiment lexicon, creating a loop from which it is difficult to infer the overall sentiment to the target entities. We propose the application of conditional random fields (CRF) to predict opinion target labels. More specifically, we exploit a set of opinion patterns to extend an opinion word lexicon and then propose to apply a CRF algorithm to detect the interactions between opinion expressions and opinion targets.

sentiment analysis, opinion target

1. INTRODUCTION

Social media has extended people's online interactions beyond simply sharing and commenting on what is happening around them, to exchanging advice and opinions with other members of the same sociosphere. This phenomenon has sparked a relationship between people's opinions and their opinion target. The information targeting the opinion targets is generally controlled by users and consumers (Jansen et al. 2009). Unlike user generated text, where the user (opinion holder) expresses freely her opinion, news articles contain a more structured text with one or more opinion holders targeting several opinion targets. This paper addresses the problem of classifying accurately the sentiment in news articles, as well as the respective sentiment target. The detection of opinion holders and targets in news articles will allow to have a better understanding of the relations between people, organizations and/or countries (Kim and Hovy (2006)). Figure 1 illustrates the opinions expressed in a news article about Internet regulations. In this example, we observe that the opinions expressed in the news article target multiple opinion targets, e.g., President Obama and U.S. data sector.



Obama's plan to regulate the Internet would do more harm than good

President Obama's call this week to regulate the Internet as a public utility is like pushing to replace the engine of a car that runs perfectly well.

The **U.S. data sector** is the envy of the world, administering a powerful boost to consumer welfare, generating high-paying jobs and encouraging tens of billions of dollars in corporate investment. (...)
Putting the Federal Communications Commission in charge of regulating broadband rates and micromanaging Web services, as the **president** proposes, would slow innovation and raise costs.

Figure 1: *Multiple opinion targets in a news article about Internet regulation.*

The analysis of opinionated text, also known as subjective text, involves the detection of words, phrases or sentences that express a sentiment. Although this area has been researched in academia, the problem is still far from being solved Liu (2012). One of the main challenges is that opinionated language varies over a broad range of discourse, and a system with a fixed vocabulary will not be enough to represent users' opinion. Another challenge is to identify relevant mentions to opinion targets which are accompanied by related sentiment words. From an algorithmic perspective, the challenge is to analyse how these sentiment words affect the public image of the opinion targets. Previous work (Hu and Liu (2004); Liu (2012)) has introduced significant advances in detecting product aspects or features, and it is reasonable to apply such methods by analysing how sentiment words affect named entities' reputation. However, unlike products, opinions about named entities are not structured around a fixed set of aspects or features, which implies a more challenging task (Albornoz et al. (2012)).

Proceedings of the 6th Symposium on
Future Directions in Information Access 2015

2. RELATED WORK

Sentiment analysis employs various techniques for detecting words that communicate a positive or negative emotion. These words are commonly known as sentiment words or opinion words. Beyond words, n-grams (contiguous sequence of n words) and idiomatic expressions are commonly used as sentiment words, such as for example “terrible”, “quite wonderful”, and “break a leg”. At document- or sentence- level, sentiment words can be used to predict sentiment classes for users opinions (Liu (2012)). Unlike sentiment analysis at document- or sentence- level, entity- or aspect-level allows for a fine-grained analysis. Entity- or aspect- level sentiment analysis captures specific product features that users dislike or like (Hu and Liu (2004)). For example, Turney (2002) proposed a document level approach to evaluate reviews polarity in which an unsupervised learning algorithm was used to evaluate review’s polarity. For each review, the authors compute the average polarity of its constituent words or phrases. Other works (Pang et al. (2002); Heerschop et al. (2011)) have addressed the sentiment analysis task by using a document-level approach. A common use of sentence-level sentiment analysis is to capture subjective sentences (Wiebe et al. (1999)). When classifying subjectivity, the goal is to distinguish between sentences that express factual information (objective) and sentences that express an opinion (subjective) (Hatzivassiloglou and Wiebe (2000)).

The task of detecting overall sentiment, opinion holders and targets implies several steps (Liu (2012)). In a sentence-level sentiment analysis approach, Meena and Prabhakar (2007) showed that rules based on atomic sentiments of individual phrases can be helpful to decided the overall sentiment of a sentence. However, in Meena at al. work, only adjectives and verbs were considered as features, which implies that only those can be related to the opinion target. Furthermore, as Wilson et al. (2009) showed, other word families (e.g., nouns) may share dependency relations with opinion targets (also referred as aspects), which might be indicative of the sentiment expressed towards those terms. In another work by Gildea and Jurafsky (2002), the authors introduced a system based on statistical classifiers to identify semantic relationships. Their system analyses the prior probabilities of various combinations of semantic roles (predicate verb, noun, or adjective) to automatically label domain-specific semantic roles such as Agent, Patient, Speaker or Topic. Similarly to the semantic roles’ detection introduced by Gildea et al., we propose to analyze sentences lexical and syntactic relations to automatically label opinion targets.

3. OPINION WORDS AND OPINION-PHRASES

We employ Moghaddam and Ester (2012) semantic relationships between words to extract opinion-phrases. These have proven to be quite successful in asserting semantic relations between opinion phrases. Table 1 shows the applied rules. For example, rules number 1 and 5 are able to extract the opinion-phrases *(works, amazing)* and *(small, blurry)* from sentences ”The automode works amazing.” and ”The LCD is small and blurry.” respectively.

Table 1: *Patterns to capture opinion-phrases (N is a noun, A is an adjective, V is a verb, h is a head term, m is a modifier, and <h, m> is an opinion phrase)*

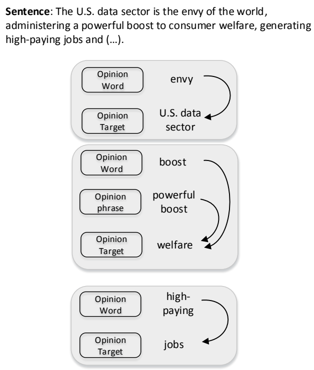
1. amod(N, A) $\rightarrow <$ N, A $>$
2. acomp(V, A) + nsubj(V, N) $\rightarrow <$ N, A $>$
3. cop(A, V) + nsubj(A, N) $\rightarrow <$ N, V $>$
4. dobj(V, N) + nsubj(V, N0) $\rightarrow <$ N, V $>$
5. $<$ h1, m $>$ + conj and (h1,h2) $\rightarrow <$ h2, m $>$
6. $<$ h, m1 $>$ + conj and(h1, h2) $\rightarrow <$ h, m2 $>$
7. $<$ h, m $>$ + neg(m, not) $\rightarrow <$ h, not + m$>$
8. $<$ h, m $>$ + nn(h, N) $\rightarrow <$ N + h, m $>$
9. $<$ h, m $>$ + nn(N, h) $\rightarrow <$ n + N, m $>$

The proximity between an opinion target and a single opinion word is key to building the opinion target semantic roles. For this reason, we have used SentiWordNet, which is a popular sentiment dictionary introduced by Esuli and Sebastiani (2006). SentiWordNet is a lexicon created semi-automatically by means of linguistic classifiers and human annotation. In SentiWordNet, each synset is annotated with its degree of positivity, negativity and neutrality.

4. THE PROPOSED MODEL

An important first step to extracting opinion targets in news articles, is understanding how an opinion word is semantically related to an opinion target. To this end, we propose a sentence-level approach, where our method will identify the opinion words and opinion phrases (Section 3). Figure 2 provides an example on how we aim to decompose each sentence.

We suggest to deal with the task of identifying opinion targets as a sequence labelling problem. The problem of opinion target extraction as a sequence labelling task using CRFs, is defined as follows. Given a sequence of tokens, $x = x_1x_2...x_n$ we need to generate a sequence of labels $y = y_1y_2...y_n$. To train the model, a set of labels are defined as ’OW’ and ’OT’, where ’OW’ corresponds to an opinion word or phrase, and ’OT’ to an opinion target. Similarly to Choi et al. (2005), opinion holders detection model, we create a linear-chain CRF



Sentence: The U.S. data sector is the envy of the world, administering a powerful boost to consumer welfare, generating high-paying jobs and (…).

Figure 2: *An overview of our opinions words extraction.*

based on an undirected graph $G = (V, E)$, where for each n tokens of a sentence V is the set of random variables $Y = \{Y_i | 1{<}i \leq n\}$. $E = \{(Y_{i-1}, Y_i) | 1{<}i \leq n\}$ is the set of $n-1$ edges forming a linear chain. According to Lafferty et al. (2001) the conditional probability of a sequence of labels y given a sequence of tokens x is given by:

$$P(y|x) = \frac{1}{Z_x} exp\Big(\sum_{i,k} \lambda_k f_k(y_{i-1}, y_i, x) + \sum_{i,k} \lambda_k^{'} f_k^{'}(y_i, x)\Big) \quad (1)$$

$$Z_x = \sum_y exp\Big(\sum_{i,k} \lambda_k f_k(y_{i-1}, y_i, x) + \sum_{i,k} \lambda_k^{'} f_k^{'}(y_i, x)\Big) \quad (2)$$

where Z_x is a normalization constant for each sentence x, $f_k(\ldots)$ is a binary feature indicator function, λ_k is the weight of each feature function, and k are the number of features defined for edges and $k^{'}$ for nodes.

5. EXPERIMENTS

5.1. Dataset

The goals are experiments first, to see how accurately we can perform a binary sentiment classification, and second, to examine the correlation between opinion phrases and opinion targets. For this analysis, one challenge to overcome is the lack of labeled data. To this end, we have selected a labeled dataset from SemEval-2014 challenge. This dataset contains opinionated sentences from the restaurants domain, and it is part of the Task 4: *Aspect Based Sentiment Analysis* of the abovementioned challenge [1]. In addition the dataset has a total of 1601 annotated sentences in which 1198 and 403 are positive and negative respectively. In addition, the dataset presents a mean of 66 characters and 12 words per sentence.

5.2. Sentiment classification

For our sentiment classification task, the sentences are classified according to a deterministic binary classification in which sentences are classified as either positive or negative. TO classify the sentences we applied a 10-fold cross validation using the Weka [2] implementation of SVM (Support Vector Machines). Table 2 shows the initial sentiment classification results.

Table 2: *Sentiment classification of comments from restaurant reviews.*

Polarity	Precision	Recall	F1
positive	0.792	0.915	0.849
negative	0.583	0.331	0.422

We note, that the classifier performs better on positive sentences. One reason for this could the imbalanced nature of the dataset. Also, as it has been demonstrated by previous work (Liu (2012)), users tend to frequently apply the same opinion word both in positive and negative contexts. This sentiment classification experiment aims to validate the quality of the selected opinion words and opinion phrases in a sentiment classification task.

5.3. Opinion phrases and opinion targets

In the present work we argue that there are many semantic relations between opinion words and opinion phrases that semantic relations analysis is not able to capture, i.e. subject and object relations. Also, as expected, we notice an intersection between opinion phrases and opinion targets. For example, in the sentence ”The service was excellent and the food was delicious.” the labeled opinion targets are ”food” and ”service” and the extracted opinion phrases are ”food delicious”, ”service excellent” and ”service delicious”. In this context, an opinion phrase is defined as a pair (aspect, opinion), therefore aspect has a high probability to be an opinion target. For the extracted opinion phrases and labeled opinion targets we observe a Jaccart similarity of 0.28. Here, Jaccart similarity refers to the quotient between the intersection of opinion- phrases and targets. Although we observe intersection between these objects there are many opinion targets that are not within the obtained opinion phrases.

5.4. Future work: Opinion targets in news articles

We observed that grammatical dependencies can be used to extract aspects and opinion phrases. However, it is noticeable that a more in-depth approach should be applied to improve the opinion

[1] http://alt.qcri.org/semeval2014/task4/

[2] http://www.cs.waikato.ac.nz/ml/weka/

targets extraction. As future work, we aim at developing the proposed CRF model to obtain a higher coverage of the opinion targets. Finally, for the experiments shown in this section, we used a dataset from the restaurants' domain. In addition, the sentences were extracted from users' reviews, which have a structure that is considerable different to that observed in news articles. We also aim at obtaining an labeled news articles dataset to extend the opinion prediction model to this domain as well.

6. CONCLUSIONS

In this paper we discussed techniques to detect opinion targets. In opinionated sentences, an opinion target is the entity that is targeted by the sentiment expressed in the sentence. Our experimental results show that opinion phrases present an clear intersection with opinion targets. However, it is evident that there are many opinion targets that are not captured by this method. We believe that this is because fixed language pattern rules are not enough to cover the range of discourse used to express an opinion, as well as the respective target. In the future, we plan to extend our work to a news articles dataset which characterized by a different type of discourse, and apply a method based on CRF to detect opinion targets language patterns.

REFERENCES

Albornoz, J., I. Chugur, and E. Amigó (2012). Using an Emotion-based Model and Sentiment Analysis Techniques to Classify Polarity for Reputation. In *Conf. and Labs of the Evaluation Forum, Online Working Notes (CLEF)*.

Choi, Y., C. Cardie, E. Riloff, and S. Patwardhan (2005). Identifying Sources of Opinions with Conditional Random Fields and Extraction Patterns. In *Proc. of the Conf. on Empirical Methods in Natural Language Processing (EMNLP)*.

Esuli, A. and F. Sebastiani (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. *Proc. of the 5th Conf. on Language Resources and Evaluation (LREC)*.

Gildea, D. and D. Jurafsky (2002). Automatic Labeling of Semantic Roles. *Computational Linguistics*.

Hatzivassiloglou, V. and J. Wiebe (2000). Effects of adjective orientation and gradability on sentence subjectivity. *Proc. of the 18th Conf. on Computational Linguistics (COLING)*.

Heerschop, B., F. Goossen, A. Hogenboom, F. Frasincar, U. Kaymak, and F. De Jong (2011). Polarity analysis of texts using discourse structure. *Proc. of the 20th Conf. on Information and Knowledge Management (CIKM)*.

Hu, M. and B. Liu (2004). Mining opinion features in customer reviews. *Proc. of the Association for the Advancement of Artificial Intelligence 19th Conf. on Artifical Intelligence*.

Jansen, B., M. Zhang, K. Sobel, and A. Chowdury (2009). Twitter Power: Tweets As Electronic Word of Mouth. *Journal of the American Society for Information Science and Technology*.

Kim, S. and E. Hovy (2006). Extracting opinions, opinion holders, and topics expressed in online news media text. *Proc. of the Workshop on Sentiment and Subjectivity in Text*.

Lafferty, J., A. McCallum, and F. Pereira (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proc. of the 18th Conf. on Machine Learning (ICML)*.

Liu, B. (2012). Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies*.

Meena, A. and T. Prabhakar (2007). Sentence Level Sentiment Analysis in the Presence of Conjuncts Using Linguistic Analysis. In *Proc. of the 29th European Conf. on Advances in Information Retrieval (ECIR)*.

Moghaddam, S. and M. Ester (2012). On the Design of LDA Models for Aspect-based Opinion Mining. *Proc. of the 21st Conf. on Information and Knowledge Management (CIKM)*.

Pang, B., L. Lee, and S. Vaithyanathan (2002). Thumbs up?: sentiment classification using machine learning techniques. *Proc. of the Conf. on Empirical Methods in Natural Language Processing (EMNLP)*.

Turney, P. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. *Proc. of the 40th Annual Meeting on Association for Computational Linguistics (ACL)*.

Wiebe, J., R. Bruce, and T. O'Hara (1999). Development and use of a gold-standard data set for subjectivity classifications. *Proc. of the 37th of the Association for Computational Linguistics on Computational Linguistics (ACL)*.

Wilson, T., J. Wiebe, and P. Hoffmann (2009). Recognizing Contextual Polarity: An Exploration of Features for Phrase-level Sentiment Analysis. *Journal Computational Linguistics*.