– Preface –
SIGIR 2015 International Workshop on Social Personalization & Search

August 10, 2015

1 Overview

For the Social Personalization & Search workshop[1], we invited researchers from all over the world working in the fields of Information Retrieval, Personalization, User Modeling, and Recommender Systems where the social dimension plays a fundamental role. The workshop is examining several approaches that leverage the social side of the search process on two main contexts: (a) using social data for improving search and recommendations, and (b) search as a social process, collaborative IR. We invited submissions that included the following topics:

- search and recommendations based on social links
- search and recommendations in collaborative tagging systems
- group-level search personalization
- search and recommendations in blogs and microblogs
- approaches for social personalization in recommender systems
- approaches on personalized social collaboration
- approaches on social linking
- methods for social search and navigation
- methods for social predictive models
- social methods for information visualization
- any other methods that exploit new forms of social data for search and personalization

The goal of this workshop was to share and discuss research that goes hopefully beyond classic personalization techniques, trying to capitalize potentially useful information available in social data for paving the way to more efficient personalized information access technologies. The workshop received nine submissions this year of which we accepted five to be presented. Each submission was carefully peer-reviewed by at least three people from the PC. In addition to this, the workshop featured two invited talks by Ricardo Baeza-Yates (Yahoo!

[1] http://socialcomputing.ing.puc.cl/sps2015/

Labs) and Paul Bennett (Microsoft). We thank the ACM SIGIR conference organizers for making this workshop possible, our program committee members, who did a great job in reviewing and discussing the contributions submitted to our workshop, as well as our two invited speakers Ricardo and Paul. Finally, also a big THANK YOU to Alejandro Bellogin who helped us with many things regarding the organization.

2 Program

2.1 Invited Talks

- *Wisdom of Crowds or Wisdom of a Few?* by Ricardo Baeza-Yates
- *Search from Personal to Social Context: Progress and Challenges* by Paul Bennett

2.2 Presentations

- *Improving Contextual Suggestions using Open Web Domain Knowledge* by Thaer Samar, Alejandro Bellogin, Arjen de Vries
- *Finding Intermediary Topics Between People of Opposing Views: A Case Study* by Eduardo Graells-Garrido, Mounia Lalmas, Ricardo Baeza-Yates
- *Analysis of Sentiment Communities in Online Networks* by Davide Feltoni Gurini, Fabio Gasparetti, Alessandro Micarelli, Giuseppe Sansonetti
- *Retrieving Relevant Conversations for Q&A on Twitter* by Jose Miguel Herrera, Denis Parra, Barbara Poblete
- *Persona-ization: Searching on Behalf of Others* by Paul Bennett, Emre Kiciman

3 Organization

3.1 Chairs

- *Christoph Trattner* is the head of the Social Computing Research Area at Know-Center, Austria's research competence center for Data-driven Business and Big Data Analytics. He holds a PhD (with hons) in CS from Graz University of Technology, Austria and he is currently enrolled as an ERCIM Alain Bensoussan fellow with NTNU, Norway. His research interests include Information Retrieval, Web Science, Data Mining and Recommender Systems, especially in the Social Context.

- *Denis Parra* is Assistant Professor at the Department of Computer Science, School of Engineering in PUC Chile. He received a PhD in Information Science from University of Pittsburgh (PA, USA, 2013) and currently conducts research on Personalization, Social Network Analysis and Information Visualization at the PUC Social Computing and Visualization (SoCVis) Lab. His research interests include Statistical Analysis, Recommender Systems and SNA.

- *Peter Brusilovsky* is currently Professor of Information Science and Intelligent Systems at the University of Pittsburgh, where he directs Personalized Adaptive Web Systems (PAWS) lab. He has been working in the field of adaptive systems, user modeling, and intelligent user interfaces for more than 20 years. He published numerous papers and edited several books on adaptive hypermedia and the adaptive Web. Peter is the Associate Editor-in-Chief of IEEE TLT and a board member of several journals including UMUAI, ACM TWEB, and Web Intelligence and Agent Systems.

- *Leandro Marinho* is currently adjunct professor at the Federal University of Campina Grande, Brazil. In 2010 he received his Ph.D. degree in computer science from the University of Hildesheim, Germany. His research interests include Machine Learning, Recommender Systems, the Semantic Web and Social Media Mining. At UFCG, he teaches Discrete Mathematics and serves as the coordinator of the undergraduate program in Computer Science.

3.2 Program Committee Members

- Luca Maria Aiello (Yahoo! Labs)
- Jussara Almeida (UFMG)
- Nazareno Andrade (Universidade Federal de Campina Grande)
- Krisztian Balog (University of Stavanger)
- Alejandro Bellogin (UAM)
- Steven Bourke (Schibsted)
- Robin Burke (DePaul University)
- Ernesto Diaz-Aviles (IBM Reseach)
- Lucas Drumond (University of Hildesheim)
- Michael Ekstrand (Texas State University)
- Alexander Felfernig (Graz University of Technology)
- Zeno Gantner (Nokia gate5 GmbH)
- Ruth Garcia-Gavilabes (Barcelona Media)
- Eduardo Graells (Telefonica I+D)
- Michael Granitzer (University of Passau)
- Ido Guy (Yahoo!)
- Eelco Herder (L3S)
- Shuguang Han (University of Pittsburgh)

- Andreas Hotho (University of Wuerzburg)
- Geert-Jan Houben (TU Delft)
- Sharon Hsiao (Arizona State University)
- Kris Jack (Mendeley)
- Alexandros Karatzoglou (Telefonica Research)
- Bart Knijnenburg (University of California)
- Milos Kravcik (RWTH Aachen University)
- Kjetil Norvag (Norwegian University of Science and Technology)
- Barbara Poblete (University of Chile)
- Giancarlo Ruffo (Universita' di Torino)
- Shaghayegh Sahebi (University of Pittsburgh)
- Alan Said (Recorded Future)
- Markus Schedl (Johannes Kepler University)
- Marc Smith (Connected Action Consulting Group)
- Nava Tintarev (University of Aberdeen)
- Eduardo Veas (Know-Center)
- Tao Ye (Pandora Inc)
- Zhen Yue (Yahoo Labs)
- Arkaitz Zubiaga (University of Warwick)

Wisdom of Crowds or Wisdom of a Few?

Ricarddo Baeza-Yates
Yahoo! Labs
Sunnyvale, USA

ABSTRACT

In this keynote we give an introduction to wisdom of crowds in the Web, the long tail of web content, and the bias involved in the generation of user generated content (UGC). This bias creates the wisdom of ad hoc crowds or the wisdom of a few. Although it is well known that user activity in most settings follows a power law, that is, few people do a lot, while most do nothing, there are few studies that characterize well this activity. In a recent analysis of social network data we corroborated that a small percentage of the active users (passive users are the majority) represent at least the 50% of the UGC. As a sub-product, we also found a lower bound for the digital desert, the content in the Web that nobody reads. These results implies that most of the wisdom comes from a few users, which is not that surprising, as the Web is a reflection of our own society, where economical or political power also is in the hands of minorities.

Keywords

Social Personalization and Search, Wisdom of the Crowds

1. BIO

Ricardo Baeza-Yates is VP of Research for Yahoo Labs leading teams in United States, Europe and Latin America since 2006 and based in Sunnyvale, California, since August 2014. During this time he has lead the labs in Barcelona and Santiago de Chile. Between 2008 and 2012 he also oversaw the Haifa lab. He is also part time Professor at the Dept. of Information and Communication Technologies of the Universitat Pompeu Fabra, in Barcelona, Spain. During 2005 he was an ICREA research professor at the same university. Until 2004 he was Professor and before founder and Director of the Center for Web Research at the Dept. of Computing Science of the University of Chile (in leave of absence until today). He obtained a Ph.D. in CS from the University of Waterloo, Canada, in 1989. Before he obtained two masters (M.Sc. CS & M.Eng. EE) and the electronics engineer degree from the University of Chile in Santiago. He is co-author of the best-seller Modern Information Retrieval textbook, published in 1999 by Addison-Wesley with a second enlarged edition in 2011, that won the ASIST 2012 Book of the Year award. He is also co-author of the 2nd edition of the Handbook of Algorithms and Data Structures, Addison-Wesley, 1991; and co-editor of Information Retrieval: Algorithms and Data Structures, Prentice-Hall, 1992, among more than 500 other publications.

From 2002 to 2004 he was elected to the board of governors of the IEEE Computer Society and in 2012 he was elected for the ACM Council. He has received the Organization of American States award for young researchers in exact sciences (1993), the Graham Medal for innovation in computing given by the University of Waterloo to distinguished ex-alumni (2007), the CLEI Latin American distinction for contributions to CS in the region (2009), and the National Award of the Chilean Association of Engineers (2010), among other distinctions. In 2003 he was the first computer scientist to be elected to the Chilean Academy of Sciences and since 2010 is a founding member of the Chilean Academy of Engineering. In 2009 he was named ACM Fellow and in 2011 IEEE Fellow.

Search from Personal to Social Context: Progress and Challenges

Paul N. Bennett
Microsoft Research
Redmond, USA

ABSTRACT

User and behavioral modeling plays a critical role in a variety of online services such as web search, advertising, e-commerce, and news recommendation. For example, our ability to accurately interpret the intent of a web search can be informed by knowledge of the web pages a searcher was viewing when initiating the search or recent actions of the searcher such as queries issued, results clicked, and pages viewed. In this talk, I will describe a recent framework for personalized search which improves the quality of search results by enabling a representation of a broad variety of context including the searcher's long-term interests, recent activity, current focus, and other user characteristics. Then, I will review a variety of related work that extends these approaches from signals focused on the individual to social signals such as likes, cohorts, and affiliation networks. Finally, I'll speculate on how social signals and networks can provide directions for relatively unexplored directions in social personalized retrieval.

Keywords

Social Personalization and Search, Social signals

1. BIO

Paul Bennett is a Senior Researcher in the Context, Learning & User Experience for Search (CLUES) group at Microsoft Research where he focuses on the development, improvement, and analysis of machine learning and data mining methods as components of real-world, large-scale adaptive systems. His research has advanced techniques for ensemble methods and the combination of information sources, calibration, consensus methods for noisy supervision labels, active learning and evaluation, supervised classification (with an emphasis on hierarchical classification) and ranking with applications to information retrieval, crowdsourcing, behavioral modeling and analysis, and personalization. His recent work has been recognized with a SIGIR 2012 Best Paper Honorable Mention and a SIGIR 2013 Best Student Paper award. He completed his dissertation on combining text classifiers using reliability indicators in 2006 at Carnegie Mellon where he was advised by Profs. Jaime Carbonell and John Lafferty.

Improving Contextual Suggestions using Open Web Domain Knowledge

Thaer Samar,[1] Alejandro Bellogín,[2] and Arjen de Vries[1]

[1] Centrum Wiskunde & Informatica, Amsterdam, The Netherlands
[2] Universidad Autónoma de Madrid, Madrid, Spain

{samar|arjen}@cwi.nl, {alejandro.bellogin}@uam.es

ABSTRACT

Contextual suggestion aims at recommending items to users given their current context, such as location-based tourist recommendations. Our contextual suggestion ranking model consists of two main components: selecting candidate suggestions and providing a ranked list of personalized suggestions. We focus on selecting appropriate suggestions from the ClueWeb12 collection using tourist domain knowledge inferred from social sites and resources available on the public Web (Open Web). Specifically, we generate two candidate subsets retrieved from the ClueWeb12 collection, one by filtering the content on mentions of the location context, and one by integrating domain knowledge derived from the Open Web. The impact of these candidate selection methods on contextual suggestion effectiveness is analyzed using the test collection constructed for the TREC Contextual Suggestion Track in 2014. Our main findings are that contextual suggestion performance on the subset created using Open Web domain knowledge is significantly better than using only geographical information. Second, using a prior probability estimated from domain knowledge leads to better suggestions and improves the performance.

1. INTRODUCTION

Recommender systems aim to help people find items of interest from a large pool of potentially interesting items. The users' preferences may change depending on their current context, such as the time of day, the device they use, or their location. Hence, those recommendations or suggestions should be tailored to the context of the user. Typically, recommender systems suggest a list of items based on users preferences. However, awareness of the importance of context as a third dimension beyond users and items has increased, for recommendation [1] and search [10] alike. The goal is to anticipate users' context without asking them. This problem – known as *contextual suggestion* in Information Retrieval (IR) and *context-aware recommendation* in the Recommender Systems (RS) community – is far from being solved. Depending on the type of context taken into account (time, location, group, short-term preferences, etc.), different techniques have been proposed. We use the definition of context stated in TREC's Contextual Suggestion (CS) track [5]: a context consists of a geographical location (a city and its corresponding state in the United States). The CS track investigates search techniques for complex information needs that are highly dependent on context and user preferences. Submission based on documents collected from either the Open Web or ClueWeb12 collection has been allowed since 2013, and the goal is to provide a list of ranked suggestions per (user, context) pair. An earlier analysis of the track's empirical results (in 2013 and 2014) has shown that runs based on the Open Web usually achieve higher effectiveness than those based on ClueWeb12 collection [6, 7].

The majority of existing studies have relied on location-based social networks from the Open Web that are specialized in providing tourist suggestions, such as Yelp and Foursquare; focusing on re-ranking the candidate suggestions based on user preferences. The main problem addressed then is to model user interests through content-based recommendation, considering evidence in the form of terms taken from the textual descriptions [12] or categories [14] of suggestions in the user profile and their associated ratings, and approaches to rank suggestions based on their similarity with the user profile. Likewise, in [8] the authors combine various user-dependent and venue-dependent features, including the aforementioned descriptions and category features, in one ranking model. However, using the ClueWeb12 collection as source of attractions requires first the selection of candidate documents, to be ranked later based on user preferences. The selection of candidate documents is a challenging task, since the (potentially) relevant suggestions have to be selected from this large collection.

In this paper, we use domain knowledge inferred from location-based social networks on the Open Web for selecting suggestions from ClueWeb12. We evaluate our contextual suggestion model on two sub-collections of the ClueWeb12 collection. One of the two sub-collections was generated using location-based social networks to annotate the candidate documents from ClueWeb12 collection. We discuss how explicit representation of knowledge about the tourism domain available on the location-based social networks improves the effectiveness of our contextual suggestion model. We show that the same contextual suggestion model for recommendation achieves an order of magnitude difference in effectiveness, depending on the approach used to derive the candidate suggestions from ClueWeb12. We address the following research questions:

RQ1 Can we improve the quality of contextual suggestions based on ClueWeb12 collection by applying domain knowledge inferred from location-based APIs?

RQ2 What is the impact of the type of domain knowledge inferred on recommendation effectiveness?

RQ3 Can we improve the results by modeling the candidate selection process probabilistically?

2. EXPERIMENTAL SETUP

2.1 Dataset and Evaluation

The models and approaches presented in this paper have been evaluated by participating in the TREC 2014 Contextual Suggestion track (CS 2014). The test dataset consists of user profiles and contexts (50 cities situated in the United states), and the task is to provide a ranked list of suggestions for each (user, context) pair. The

user profiles were constructed based on the training data, which consists of 100 example suggestions located in two cities, `Chicago, IL` and `Santa Fe, NM`. Each user profile represents the rating given by a crowd-source user to the examples. It consists of two ratings per suggestion, on a 5-point scale; one rating for a suggestion's description (i.e., a snippet), and another rating for its actual content (i.e., once the web page has been visited). In total, 299 (user, context) pairs have been judged. For these pairs, the top-5 documents of every submission have been judged by the assessors (profile owners). Judgments range from 0 (strongly uninterested) to 4 (strongly interested). In order to judge the geographical relevance of the suggestion, assessors were asked to judge whether the suggestion is located in the city it was suggested for. In addition to the crowd-source users, geographical relevance was also judged by NIST assessors. In both cases the geographical judgment ranges from 0 (not geographically appropriate) to 2 (geographically appropriate). Since submissions were allowed to be either from the Open Web or the ClueWeb12 collection, in the relevance judgments suggestions from the Open Web were identified by their URLs, while suggestions from ClueWeb12 collection were identified by their ClueWeb12 ids. For evaluating the performance of submitted runs, Precision@5 (P@5), Mean Reciprocal Rank (MRR), and a modified Time-Biased Gain (TBG) [4] were used as the "official TREC metrics". These metrics consider geographical and profile relevance (both in terms of document and description judgments), taking as thresholds a value of 1 and 3 (inclusive), respectively.

Our initial analysis is based on the two runs that our team submitted for evaluation. Both runs are based on sub-collections of candidate suggestions belonging to the ClueWeb12 collection; the first using the **GeographicFiltered** sub-collection that we describe in Section 3.3.1 and the second one using the **TouristFiltered** sub-collection described in Section 3.3.2. In our analyses, we refer to these runs by the name of the sub-collection that it is based on.

2.2 URL Normalization

A recurring pre-processing step to produce the various results reported in the paper concerns the normalization of URLs. We have normalized URLs consistently by removing their `www, http://, https://` prefixes, as well as their trailing "forwarding slash" character `/`, if any. In the special case of the URL referencing an `index.html` web page, the `index.html` string is stripped from the URL before the other normalizations are applied.

3. CONTEXTUAL SUGGESTION MODEL

In this section, we formulate the problem and describe a general framework for finding and providing personalized recommendations based on user preferences. Then, we describe the two main components of our model. The first component represents our approach for generating personalized ranked suggestions to the user based on her preferences (Section 3.2). The second component describes our approach for modeling the selection of candidates from ClueWeb12 collection (Section 3.3).

3.1 General Model and Problem Formulation

We assume that we have a set of suggestions – represented by a URL and a description – that have been judged by a set of users. The goal is to provide a ranked list of personalized suggestions for the users in new contexts. We exploit the user preferences and the given suggestion descriptions to model a textual user's positive and negative profiles into a similarity ranking model that is able to regulate the impact of the positive and negative profiles to generate a final scoring. We adopt a standard approach to content-based recommendation to determine a ranked list of suggestions:

$$P_{rel}(u, s) = P(s) \cdot SIM(u, s) \quad (1)$$

$P(s)$ is a probability that estimates how likely it is that suggestion s is relevant to the task, and controls the suggestions considered. We have experimented with different approaches to estimate this probability, described in detail in Section 3.3. Note that $P(s)$ does not necessarily depend on the user (the equivalent to the queries in traditional retrieval models), although it may depend on the context; it can be compared to the "prior probability of relevance" of traditional information retrieval models. If the range of $P(s)$ is restricted to discrete values 0 and 1, then $P(s)$ acts as a Boolean filter that selects candidate suggestions based on some features.

3.2 Personalization

Similarity function $SIM(u, s)$ represents the (content-based) similarity between user interests and candidate suggestions, and determines the personalization of recommendations to the user's interests. We follow an approach to modeling user preferences that has been used widely in the literature on contextual suggestion; consider for example [2, 11, 12]. Descriptions of the previously rated attractions provide the basis to construct two user profiles for each user. The positive profile u^+ represents the attractions that the user u likes, whereas the negative profile u^- represents the attractions that the user u dislikes. We use the value 2.5 (since ratings are on 0 to 4 scale) as a threshold to discriminate between liked and disliked attractions. We compute the similarity score between a candidate suggestion s and a user u as follows:

$$SIM(u, s) = \lambda \cdot SIM(u^+, s) - (1 - \lambda) \cdot SIM(u^-, s) \quad (2)$$

where $SIM(u^+, s)$ is the similarity between user's positive profile and the candidate document, while $SIM(u^-, s)$ is the similarity between user's negative profile and the candidate document. λ is the parameter that regulates the contribution of the $SIM(u^+, s)$ and $SIM(u^-, s)$ to the final score. We used 5-fold cross-validation on training data to find the optimal $\lambda = 0.7$, which was selected from $[0, 1]$ in 0.1 steps. For this experiment, we considered the cosine similarity (based on term frequencies). This has been done after transforming the suggestions and the user profiles from text-representation into a weighted vector-based representation. In this transformation, we filter out the HTML tags from the content of the documents, apply common IR parsing techniques including stemming and stop-word removal.

3.3 Selection Methods of Candidates

The selection of candidate suggestions plays an important role for providing good suggestions to the users. We have already presented how previous works address the contextual suggestion challenge by using a variety of public tourist APIs – including Google Places, WikiTravel, Yelp, and Foursquare – to obtain a set of suggestions. Queries issued are usually related to the target context (location), either given by its name (i.e., *Chicago, IL*) or its latitude and longitude coordinates (i.e., (41.85003, −87.65005)). Collecting suggestions from the ClueWeb12 collection poses however new challenges, different from "just" constructing the right query to issue at location-based web services. We formulate the problem of candidate selection from ClueWeb12 as follows. We have a set of contexts (locations) `C` – which correspond to US cities – provided by the `CS` track organizers. For each context $c \in C$, we generate a set of suggestions S_c from the ClueWeb12 collection, which are expected to be located in that context. We investigate two different approaches toward generating S_c. The first approach is to apply a straightforward geographical filter, based on the content

of the ClueWeb12 documents. In the second approach, we exploit knowledge derived from external resources available on the Open Web about sites that provide touristic information, and apply this knowledge to ClueWeb12 collection.

3.3.1 Geographically Filtered Sub-collection

Our main hypothesis in this approach is that a good suggestion (a venue) will contain its location correctly mentioned in its textual content. Therefore, we implemented a content-based geographical filter (named `geo_filter`) that selects documents mentioning a specific context with the format `(City, ST)`, ignoring those mentioning the city with different states or those matching multiple contexts. With this selection method we aim to ensure that the specific target context is mentioned in the filtered documents (hence, being geographically relevant documents). The documents that pass this filter form sub-collection, **GeographicFiltered**. In Equation (1), we express this geographic filtering process through probability $P(s)$, which defines the probability of a ClueWeb12 document to be a candidate suggestion. In the simplest instantiation of our model, the probability of any document in ClueWeb12 to be included in the **GeographicFiltered** sub-collection is assigned to 0 or 1 depending on whether it passes the geo_filter:

$$P(s) = \begin{cases} 1, & \text{if } (s) \text{ passes geo\_filter} \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

Approximately 9 million documents $(8,883,068)$ from the ClueWeb12 collection pass this filter.

3.3.2 Applying Domain Knowledge to Sub-collection

The sub-collection described in Section 3.3.1 only takes the context into account, however, users are not equally satisfied by any type of document when receiving contextual suggestions: they expect those documents to be *entertaining* [4]. This implies that documents about restaurants, museums, or zoos are more likely to be relevant than stores or travel agencies [11]. We incorporate this information into our sub-collection creation process by sampling from the ClueWeb12 collection considering knowledge from the tourist domain. In the following, we present alternative ways to select candidate documents from ClueWeb12 collection using different filters. Each filter represents a domain knowledge about tourist information inferred from the Open Web.

Domain-Oriented Filter.

The first type of domain knowledge depends on a list of `hosts` that are well-known to provide tourist information, and are publicly available. We manually selected the hosts $\mathcal{H}$:= {`yelp, tripadvisor, wikitravel, zagat, xpedia, orbitz,` and `travel.yahoo`}. We consider these hosts as a domain filter to select suggestions from ClueWeb12 collection. The probability of a document in ClueWeb12 to be a candidate is either 0 or 1 depending only on its host. We define the probability $P(s)$ as:

$$P(s) = \begin{cases} 1, & \text{if host}(s) \in \mathcal{H} \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

We refer to the set of documents that pass the domain filter defined in Equation (4) as *TouristSites*.

We assume pages about tourist information also have links to other interesting related pages, acknowledging the fact that pages on the same topic are connected to each other [3]. In order to maximize the extracted number of documents from the tourist domain we also consider the outlinks of documents from touristic sites. For each suggestion $s \in$*TouristSites*, we extract its outlinks `outlinks(s)` and combine all of them together in a set $\mathcal{O}$; including links between documents from two different hosts (`external links`) as well as links between pages from the same host (`internal links`). Notice that some of the outlinks may also be part of the *TouristSites* set, because of satisfying Equation (4). Next, we extract any document from ClueWeb12 whose normalized URL matches one of the outlinks in $\mathcal{O}$. The probability of document s to be selected in this case is defined as:

$$P(s) = \begin{cases} 1, & \text{if URL}(s) \in \mathcal{O} \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

The set of candidate suggestions that pass this filter is called *TouristSitesOutlinks*.

Table 1: Number of documents for each part of the **TouristFiltered** subcollection.

Filter	Number of documents
TouristSites	175,260
TouristSitesOutlinks	97,678
Attractions	102,604
TouristFiltered	375,542

Attraction-Oriented Filter.

We will now consider a different type of domain knowledge, by leveraging the information available on the Foursquare API [1]. For each context $c \in C$, we obtain a set of URLs by querying Foursquare API. If the document's URL is not returned by Foursquare, we use the combination of document name and context to issue a query to the Google search API e.g., "`Gannon University Erie, PA`" for name *Gannon University* and context *Erie, PA*. Extracting the hosts of the URLs obtained results in a set of $1,454$ unique hosts. We then select all web pages in ClueWeb12 from these hosts as the candidate suggestions, with its probability defined in the same way as in Equation 4. The set of documents that pass the host filter is referred to by *Attractions*.

Together, the three subsets of candidate suggestions *TouristSites*, *TouristSitesOutlinks* and *Attractions* form our second ClueWeb12 sub-collection that we refer to as **TouristFiltered**.

TouristFiltered := *TouristSites*∪*TouristSitesOutlinks*∪*Attractions*

Table 1 shows statistics about the documents that pass each filter.

3.3.3 Candidates Selection Prior Probability

In Sections 3.3.1 and 3.3.2, we introduced probabilities, used as binary filters so far, to decide which documents from the ClueWeb12 collection should be selected as candidates. Each of these filters represents a different kind of knowledge related to tourism inferred from the Open Web. Now, we introduce three different methods to estimate prior $P(s)$ from the **TouristFiltered** sub-collection. Two non content-based priors exploit the correlation between relevance judgments, the depth of URLs, and the filters based on location-based social networks. The third prior is based on the content of the documents found by the best location-based filter. We evaluate the effect of these different estimations $P(s) = P_s^i$, where $i \in \{1, 2, 3\}$, by applying our contextual suggestion model on the **GeographicFiltered** sub-collection.

Previous research has shown that correlations between relevance and non content-based features such as document length can be exploited to improve retrieval results, e.g. [13]. Similarly, the authors

[1] `https://developer.foursquare.com/docs/venues/search`

of [9] presented a general model of embedding non content-based features of web pages (document length, in-link count, and URL depth) as a prior probability in the ranking model. By studying the correlation between the URL depth and the relevance of the webpage, they observed that the probability of being a home page is inversely related to URL depth. Motivated by these studies, we carry out a similar analysis on the URLs of ClueWeb12 documents and the URLs of documents in the `CS` track ground truth. We use the number of slashes in the *normalized* URL to find the depth; a more fine-grained analysis like the four categories used in [9] is deferred to future work. Table 2 shows the depth distribution of URLs in the ClueWeb12 collection. We estimate the relationship between URL depth and the prior probability of relevance by analyzing the ground truth of the Open Web qrels, the ClueWeb12 qrels, as well as the URLs in the Open Web qrels that also exist in the ClueWeb12 collection. We observe in Table 3 that approximately 72% of the documents in the Open Web qrels exist at the top levels of a website (depth zero and one), and that 75% of these are relevant, consistent with findings reported in the literature; we also find that the probability of a document being relevant is inversely related to the URL depth. However, the distribution of URL depth and their corresponding relevance is different for the ClueWeb12 qrels, where the highest percentage of webpages presented (and relevant) in those runs are at depth two, one, and three (in that order).

We can now estimate a prior probability of relevance at each URL depth by combining the statistics derived from the qrels (based on the correlation between URL depth and relevance of the ClueWeb12 ground truth information presented in Table 3 with the URL depth distribution of the complete collection, Table 2):

$$P_s^1 = P_s(depth) = P(rel|URL(depth = d_i)) = \frac{c(Rel, d_i)}{c(d_i)} \quad (6)$$

Similar to how we derive a prior probability of relevance from the URL depth data, we may also use the number of relevant documents generated by each subset filter to inform the prior probability of relevance. In this case, the probability of a document to be relevant considering that it has passed a filter is defined as follows:

$$P_s^2 = P_s(filter) = P(rel|filter_i) = \frac{c(Rel, filter_i)}{c(filter_i)} \quad (7)$$

Here, we use the statistics shown in Table 1 for the total number of documents that pass each **TouristFiltered** subset filter, to normalize the total number of relevant documents in each filter. The outcome is a filter-specific approach to estimate the prior probability of relevance. A document in **GeographicFiltered** sub-collection will get the prior probability of the filter that it passes, and the maximum prior is considered if multiple filters are satisfied. For the rest of the documents in **GeographicFiltered** sub-collection that do not satisfy any filter, they will get a prior estimated by the number of relevant documents in **GeographicFiltered** sub-collection normalized by its total number of documents.

The third prior P_s^3 is a content-based derived prior, where we use a language model constructed from documents that pass the best filter in terms of highest performance values. Specifically, we learn from the documents that pass the *Attractions* filter which were part of the **TouristFiltered** run to compute the prior probabilities. The goal is to boost documents from **GeographicFiltered** sub-collection that are similar to the attraction documents. We construct two different language models. The first is from documents that pass the *Attractions* filter and were judged as relevant. The second is from documents that pass the *Attractions* filter and were judged as not relevant. After that, both sets are processed in a similar way to generate a language model: first the stop words and non-alphabetic words are removed; then, terms are ranked based on their relative frequency in each set.

Table 2: Distribution of ClueWeb12 documents over URLs depth.

Depth	count	%
0	3,726,692	0.5
1	152,584,686	21.0
2	253,913,644	35.0
3	172,258,009	23.7
4	83,629,521	11.5
5	35,464,476	4.9
6	13,495,362	1.9
7	6,756,976	0.9
8	3,693,477	0.5
11	809,692	0.1

Table 4: Performance of **GeographicFiltered** and **TouristFiltered** runs. Analysis per relevance dimension is considered; description (desc), document (doc), and geographical (geo) relevance. We denote with (all) when desc, doc, and geo relevance are considered.

Metric	**GeographicFiltered**	**TouristFiltered**
P@5_all	0.0431	0.1374
P@5_desc-doc	0.2081	0.2222
P@5_desc	0.2828	0.2788
P@5_doc	0.2620	0.2949
P@5_geo	0.1549	0.4808
TBG	0.1234	0.5953
TBG_doc	0.1287	0.6379

4. RESULTS AND ANALYSIS

We now present how we have addressed the three research questions mentioned at the beginning of the paper and the results obtained in each situation. The measures are averaged after running a 5-fold cross-validation.

4.1 Effect of Using External Domain Knowledge for Candidate Selection

In this section we study RQ1: Can we improve the quality of contextual suggestions based on ClueWeb12 collection by applying domain knowledge inferred from location-based APIs? We compare the performance of our contextual suggestion model (see Section 3) used to rank suggestions from the two presented sub-collections **GeographicFiltered** and **TouristFiltered**. We show empirically that the additional information acquired from location-based social networks provides the evidence needed to generate high quality contextual suggestions.

Table 4 summarizes the results from the evaluation, where we are initially only interested in the entries that take all relevance criteria into account, labeled by suffix `_all`. Clearly, the effectiveness using the **TouristFiltered** sub-collection outperforms the **GeographicFiltered** results by a large margin. Also, among the results obtained for the runs submitted in TREC 2014, the former approach was superior to all other submitted ClueWeb12 runs, while the latter ranked near the bottom [6]. We should emphasize that the actual method that ranks the documents is exactly the same in both cases (Section 3.2), and hence, the difference in performance should be attributed to the differences in the candidate suggestions.

We inspect the evaluation outcomes in more detail, by considering relevance dimensions individually. Recall that assessments are made considering geographical and profile relevance independently. For the latter one, the user assessed both the document and the description provided by the method. Considering this information, we recomputed the evaluation metrics while taking into account the geographical relevance provided by the assessors, as well

Table 3: Distribution of URLs depth over the documents in Open Web qrels, documents from Open Web qrels that exist in ClueWeb12 collection, and the ClueWeb12 qrels.

	Open Web runs					overlap					ClueWeb12 runs			
	All		Relevant			All		Relevant			All		Relevant	
depth	count	%	count	%	depth	count	%	count	%	depth	count	%	count	%
0	23,657	66.31	9,271	67.69	0	8,847	87.78	1,891	81.54	0	159	1.79	22	2.53
1	2,113	5.92	636	4.64	1	473	4.69	180	7.76	1	1,856	20.89	208	23.88
2	6,957	19.50	2,758	20.14	2	423	4.20	149	6.43	2	4,537	51.06	479	54.99
3	2,211	6.20	853	6.23	3	210	2.08	52	2.24	3	1,412	15.89	86	9.87
4	434	1.22	113	0.82	4	78	0.77	19	0.82	4	688	7.74	57	6.54
5	179	0.50	47	0.34	5	36	0.36	17	0.73	5	168	1.89	13	1.49
6	52	0.15	5	0.04	6	11	0.11	3	0.13	6	43	0.48	3	0.34
7	61	0.17	6	0.04	7	1	0.01	0	0.00	7	9	0.10	1	0.11
8	14	0.04	8	0.06	8	13	0.13	8	0.34	10	9	0.10	2	0.23
11	1	0.00	13,697			10,079		2,319		13	4	0.05	871	
	35,679										8,885			

Table 5: Effect of domain knowledge filters on **TouristFiltered** run performance. Union means adding suggestions from the subset filter shown in column header of current column to the previous one. The percentage shows the relative improvement in effectiveness due to filter.

	TouristSites	∪ *TouristSitesOutlinks*		∪ *Attractions*		*Attractions*
Metrics	score	score	%	score	%	score
P@5_all	0.0392	0.0518	32.1	0.1374	165.3	0.1057
P@5_desc	0.0917	0.1200	30.9	0.2788	132.3	0.1973
P@5_doc	0.1008	0.1310	30.0	0.2949	125.1	0.2101
P@5_geo	0.2067	0.2659	28.6	0.4808	80.8	0.4667

as the description and document judgments, both separately and combined (that is, a document that is relevant both based on the description and when the assessor visited its URL, denoted with prefix *desc-doc*). Table 4 shows the analysis for each relevance dimension – note that the geographical, description, and document relevance assessments affect in the same way the evaluation metrics. When all the dimensions are considered (*all* prefix), the **TouristFiltered** sub-collection is significantly better than the **GeographicFiltered** one. However, the difference in the performance between the two sub-collections decreases when we look at the relevance of a document and its description, that is, when we ignore the geographical aspect of the relevance. This means that both sub-collections are similar in terms of their appropriateness to the users, where we only consider suitability with respect to the user's profile. At the same time, we observe that the **TouristFiltered** sub-collection is more geographically appropriate, implying that using the domain knowledge to select the candidates improves the performance in that dimension. A similar observation is found when looking at the best relevance dimension: for the **GeographicFiltered** sub-collection, the best performing dimension is the document description, whereas for the **TouristFiltered** sub-collection this is the geographical aspect.

4.2 Impact of Domain Knowledge Filters

In this section, we investigate RQ2: What is the impact of the type of domain knowledge inferred on recommendation effectiveness? We provide a deeper insight on why the domain knowledge-based sub-collection improves so much over the other sub-collection on the different relevance dimensions. Table 5 presents the contribution to the relevance dimensions of each of the **TouristFiltered** sub-collection subsets, where each subset was selected based on a different domain knowledge filter.

Table 6: Effect of using a prior-probability of relevance on the **GeographicFiltered** run performance. *no prior* means applying the general ranking model with $P(s) = 1$ for documents that pass the geo_filter.

Metrics	no prior	depth prior	filter prior
P@5_all	0.0431	0.0660	0.1300
P@5_desc-doc	0.2081	0.1024	0.1912
P@5_desc	0.2828	0.1273	0.2350
P@5_doc	0.2620	0.1468	0.2579
P@5_geo	0.1549	0.3515	0.4842
TBG	0.1234	0.3007	0.5574
TBG_doc	0.1287	0.3281	0.5988

We start modifying the run based on the **TouristFiltered** sub-collection by computing effectiveness based only on suggestions from the *TouristSites* subset (second column), then we add to them suggestions from *TouristSitesOutlinks*, and finally suggestions from *Attractions* are added. The main conclusion drawn from this table is that the larger improvement in performance occurs after adding the candidates from *Attractions* subset. It is interesting to note that the performance of this part alone (last column) is comparable to that of the whole sub-collection.

4.3 Effect of Prior Probability

In this section, we investigate RQ3: Can we improve the results by modeling the candidate selection process probabilistically? In this section, we investigate the effect of adding a prior probability that we discussed in Section 3.3.3 on the performance of the contextual suggestion model. Table 6 shows the effect of depth prior, and the effect of the filter prior when applying the contextual suggestion model on the **GeographicFiltered** sub-collection. As shown in this table, there is a significant improvement on the performance of the **GeographicFiltered** sub-collection after applying the two priors independently. We observe that the domain filter prior has more impact on the performance.

Next, we study the effect of the third prior, which is a content-based derived prior, where we use a language model constructed from documents that pass the *Attractions* filter which were part of the **TouristFiltered** run. We experimented with different cut-offs for selecting the top words to form the language model, precisely the top 500, $1,000$, and $5,000$ words. Without finding a clear relation between cutoff and performance, we present results based on the top $1,000$ terms. Table 7 shows the effect of using the similarity between the language models and the **GeographicFiltered** documents as prior. We observe that the performance is worse than without a prior (compare with first column of Table 6). However,

this can also be explained by analyzing the number of documents that have judgments in the rankings generated by each method. We therefore reported also the percentage of judged documents in top-5 as well as the percentage of relevant documents among the judged, and the precision@5 with a condition that the document is judged. We now conclude that the language model generated from the relevant documents improves the performance.

Table 7: Language model constructed from relevant and not relevant documents.

Metrics	$\neg rel$	rel
P@5_all	0.0034	0.0067
P@5_doc	0.0444	0.0694
%judged@5	28.55	46.73
%rel of judged@5	38.18	54.75
P@5_doc(judged)	0.2185	0.4824

5. CONCLUSION

We have presented an approach for improving contextual suggestions based on ClueWeb12 collection. Our approach focused on selecting candidate documents from a large Web crawl (`ClueWeb12`), using tourist domain knowledge inferred from the location-based social networks from the Open Web. First, we presented Boolean filters for modeling selection of candidate suggestions, where each filter represents a different type of knowledge about the tourist domain. The filter is then integrated in the ranking model via a prior probability of relevance. Our empirical evaluation shows that using domain knowledge drawn from location-based social networks improves the performance of the contextual suggestion model when compared to the performance of the same ranking model, using the **GeographicFiltered** sub-collection that is created without any domain knowledge. Second, we found that the two sub-collections have different correlations with the dimensions of relevance considered in the evaluation (geographical and profile relevance), which opens up to investigate more the relation between the filters and the relevance dimension. Third, our analysis shows that filters used to create the **TouristFiltered** sub-collection vary in impact on contextual suggestion effectiveness. We exploit the knowledge of each filter to estimate a probability prior embedded in the ranking model using 5-fold cross-validation analysis. We also consider the correlation between URL depth of the document and its relevance, as an alternative prior. The results of this analysis on the **GeographicFiltered** sub-collection suggest that both priors improved the performance. The domain filter prior has more influence on the performance, suggesting that the domain knowledge filter captures relevance better than the depth prior. In the future, we aim to investigate the effect of the filter prior by incorporating different sources of information, such as the relation between the filter criteria and URL depth, and the relation between filter criteria and the individual dimensions of relevance.

References

[1] G. Adomavicius and A. Tuzhilin. Context-aware recommender systems. In F. Ricci, L. Rokach, B. Shapira, P. B. Kantor, F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, editors, *Recommender Systems Handbook*, chapter 6, pages 217–253. Springer, Boston, MA, 2011.

[2] M.-D. Albakour, R. Deveaud, C. Macdonald, and I. Ounis. Diversifying contextual suggestions from location-based social networks. In *Proceedings of the 5th Information Interaction in Context Symposium*, IIiX '14, pages 125–134, New York, NY, USA, 2014. ACM.

[3] B. D. Davison. Topical locality in the web. In *In Proceedings of the 23rd Annual International Conference on Research and Development in Information Retrieval (SIGIR 2000*, pages 272–279. ACM Press, 2000.

[4] A. Dean-Hall, C. L. A. Clarke, J. Kamps, and P. Thomas. Evaluating contextual suggestion. In *Proceedings of the Fifth International Workshop on Evaluating Information Access (EVIA 2013)*, 2013.

[5] A. Dean-Hall, C. L. A. Clarke, J. Kamps, P. Thomas, and E. M. Voorhees. Overview of the TREC 2012 contextual suggestion track. In E. M. Voorhees and L. P. Buckland, editors, *Proceedings of The Twenty-First Text REtrieval Conference, TREC 2012, Gaithersburg, Maryland, USA, November 6-9, 2012*, volume Special Publication 500-298. National Institute of Standards and Technology (NIST), 2012.

[6] A. Dean-Hall, C. L. A. Clarke, J. Kamps, P. Thomas, and E. M. Voorhees. Overview of the TREC 2014 contextual suggestion track. In E. M. Voorhees and A. Ellis, editors, *Proceedings of The Twenty-Third Text REtrieval Conference, TREC 2014, Gaithersburg, Maryland, USA, November 19-21, 2014*, volume Special Publication 500-308. National Institute of Standards and Technology (NIST), 2014.

[7] A. Dean-Hall, C. L. A. Clarke, N. Simone, J. Kamps, P. Thomas, and E. M. Voorhees. Overview of the TREC 2013 contextual suggestion track. In E. M. Voorhees, editor, *Proceedings of The Twenty-Second Text REtrieval Conference, TREC 2013, Gaithersburg, Maryland, USA, November 19-22, 2013*, volume Special Publication 500-302. National Institute of Standards and Technology (NIST), 2013.

[8] R. Deveaud, M.-D. Albakour, C. Macdonald, and I. Ounis. On the importance of venue-dependent features for learning to rank contextual suggestions. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pages 1827–1830, New York, NY, USA, 2014. ACM.

[9] W. Kraaij, T. Westerveld, and D. Hiemstra. The importance of prior probabilities for entry page search. In *SIGIR 2002: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 11-15, 2002, Tampere, Finland*, pages 27–34. ACM, 2002.

[10] M. Melucci. Contextual search: A computational framework. *Foundations and Trends in Information Retrieval*, 6(4-5):257–405, 2012.

[11] A. Rikitianskii, M. Harvey, and F. Crestani. A personalised recommendation system for context-aware suggestions. In *Advances in Information Retrieval - 36th European Conference on IR Research, ECIR 2014, Amsterdam, The Netherlands, April 13-16, 2014. Proceedings*, pages 63–74, 2014.

[12] M. Sappelli, S. Verberne, and W. Kraaij. Recommending personalized touristic sights using google places. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 781–784, New York, NY, USA, 2013. ACM.

[13] A. Singhal, C. Buckley, and M. Mitra. Pivoted document length normalization. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '96, pages 21–29, New York, NY, USA, 1996. ACM.

[14] P. Yang and H. Fang. Opinion-based user profile modeling for contextual suggestions. In *Proceedings of the 2013 Conference on the Theory of Information Retrieval*, ICTIR '13, pages 18:80–18:83, New York, NY, USA, 2013. ACM.

Finding Intermediary Topics Between People of Opposing Views: A Case Study

Eduardo Graells-Garrido*
Telefónica I+D
Santiago, Chile

Mounia Lalmas
Yahoo Labs
London, UK

Ricardo Baeza-Yates
Yahoo Labs
Sunnyvale, USA

ABSTRACT

In micro-blogging platforms, people can connect with others and have conversations on a wide variety of topics. However, because of homophily and selective exposure, users tend to connect with like-minded people and only read agreeable information. Motivated by this scenario, in this paper we study the diversity of intermediary topics, which are latent topics estimated from user generated content. These topics can be used as features in recommender systems aimed at introducing people of diverse political viewpoints. We conducted a case study on Twitter, considering the debate about a sensitive issue in Chile, where we quantified homophilic behavior in terms of political discussion and then we evaluated the diversity of intermediary topics in terms of political stances of users.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Information networks*

Keywords

Social Networks; Topic Modeling; Homophily; Political Diversity.

1. INTRODUCTION

Social research has shown that, while everyone indeed has a voice, people tend to listen and connect only to those of similar beliefs in political and ideological issues, a cognitive bias known as homophily [19]. This bias is present in many situations, and it can be beneficial, as communication with culturally alike people is easier to handle. However, the consequences of homophily in ideological issues are prominent, both off- and on-line. On one hand, groups of like-minded users tend to disconnect from other groups, polarizing group views. On the other hand, Web platforms recommend and adapt content based on interaction and network data of users, *i. e.*, who is connected to them and what they have liked before. Because algorithms want to maximize user engagement, they recommend content that reinforces the homophily in behavior and display only agreeable information. Such biased reinforcement, in turn, makes computer systems to recommend even more polarizing content, confining users to *filter bubbles* [20].

One way to improve the current situation is to motivate users to read challenging information, or to motivate a change in behavior through recommender systems. However, this "direct" approach has not been effective as users do not seem to value political diversity or do not feel satisfied with it [1], a result explained by *cognitive dissonance* [14], a state of discomfort that affects persons confronted with conflicting ideas, beliefs, values or emotional reactions. Conversely, Graells-Garrido *et al.* [16] proposed an "indirect" approach, by taking advantage of partial homophily to suggest similar people, where similarity is estimated according to *intermediary topics*. Intermediary topics are defined as non-conflictive shared interests between users, *i. e.*, interests where two persons of opposing views on sensitive issues could communicate and discuss without facing challenging information in a first encounter. According to the primacy effect in impression formation [2], first impressions matter, making such intermediary topics important when introducing people. In recommender systems, recommendations based on intermediary topics would indirectly address the problem of exposing people to others of opposing views in a non-challenging context.

In this work, we extend the definition of intermediary topics [16]. In addition, we formally evaluate this redefinition by considering the following research question: *are intermediary topics more diverse in terms of political stances and language than non-intermediary topics?* We approach this question by performing a case study on the micro-blogging platform Twitter, with users who discussed sensitive issues, *i. e.*, ideological or political themes that would make people reject connecting or interacting with others. In particular we focus on the analysis of discussion around *abortion* in Chile. Chile has one of the strictest abortion laws in the world [24], yet at the same time a majority of population is in favor of its legalization [10], making it a controversial topic suitable for analysis. Our contributions include a quantification of the homophilic structure of discussion around this topic in Chile, and a confirmation of the diversity of people with respect to political stances in intermediary topics.

This paper is organized as follows: after reviewing the background work (§ 2), we define the methods and concepts needed to study intermediary topics (§ 3). Then, we perform a case study in Chile (§ 4). Finally, we discuss results and implications (§ 5).

2. BACKGROUND

Homophily is the tendency to form ties with similar others, where similarity is bound to many factors, from sociodemographic to behavioral and intra-personal ones (see a literature review by McPherson *et al.* [19]). In micro-blogging platforms, homophily has been observed in terms of political leaning [5]. Because of homophily, ego-network structures can help to recommend people to interact with [11, 17].

In our work, we propose *intermediary topics* as a feature to consider when recommending users to follow. The intuition behind intermediary topics is that they focus on homophily in specific shared interests that are non confronting nor challenging, *i. e.*, unlikely to provoke cognitive dissonance. Our definition of intermediary

*Corresponding author: eduardo.graells@telefonica.com. Work carried out while the first author was a PhD student at Universitat Pompeu Fabra, Barcelona, Spain.

topics is based on topic modeling using *Latent Dirichlet Allocation* [6, 23]. In particular, we build a *topic graph* of relations between latent topics, and find which ones are more likely to include people from diverse political backgrounds by estimating the information centrality [8] of latent topics.

Although topic modeling has been used before to measure homophily by considering user similarity [26], we measure its presence as the deviation from the expected interaction behavior given the population distribution in terms of user stances on specific controversial political issues. This distinction is important given that homophily also appears in other dimensions (*e. g.*, demography).

To study political leaning in social media, in particular in microblogging platforms, the first challenge is to actually detect what is the political leaning of users, as this attribute is not usually part of a public profile. One way to address the issue of classifying users is through supervised machine learning [13] and bayesian estimation [7], among other methods. Features used in classification include vocabulary, hashtags, and connectivity with accounts with known political leaning. Knowing political alignment of users allows to study group polarization. In a work related to our case study, Yardi *et al.* [27] studied debates about abortion in Twitter, in particular between users of *pro-life* and *pro-choice* stances. Their results indicate that the interaction between users having the same stance reinforced group identity, and discussions with members of the opposite group were found to be not meaningful, partly because the interface did not help in that aspect. In our work, we focus on a previously unexplored context: a politically centralized Latin-American country [15]. We complement previous work and help to understand the differences in political discussion around the globe.

3. METHODS

In this section we present our methodology to model users' intermediary topics, which extends previous work [16].

Sensitive Issues and Shared Interests. *Sensitive issues* are political or ideological topics for which their stances or opinions tend to divide people. This considers topics like *global warming*, *social security*, *health care reforms*, and *abortion*. Such topics tend to polarize people, *i. e.*, users who support one stance in abortion do not interact with users who support another stance, a behavior explained by homophily and cognitive dissonance. Conversely, *shared interests* are topics for which their stances or opinions do not, in normal conditions, tend to divide people. As example, people who support the soccer team *F.C. Barcelona* have a rivalry with people who support *Real Madrid F.C.*, however, the selective exposure mechanism would not be activated when discriminating information coming from people who support the opposite team–in fact, in some cases, they might be interested in such information. Other contexts can be less challenging as there might be no explicit rivalries. For instance, people with different musical tastes might be interested in discussing the particularities of their preferred music styles for comparison with others. As such, those shared interests could be good features to consider when introducing people [16], specially when considering first impressions [2].

Representation of User Stances in Sensitive Issues. An assumption we make with respect to user stances is that they are linked by partisan political ideology, *e. g.*, conservative/liberal people share views on different sensitive issues. Then, to estimate user stances, we first need to be able to estimate what users say with respect to sensitive issues. In Twitter, often users annotate their tweets with *hashtags*, which are text identifiers that start with the character #. For instance, *#prochoice* and *#prolife* are two hashtags related to two abortion stances, and each one of those stances has specific words related to them (*e. g.*, *"right to choose"* is pro-choice, and *"it is life since conception"* is pro-life). Pennacchiotti *et al.* [21] call those related words *prototypical words and hashtags*. We refer to both as prototypical keywords indistinctively. For any sensitive issue under consideration, we collect relevant tweets based on prototypical keywords (*e. g.*, *#prochoice*, *#prolife*, *abortion*, *pregnancy*, *interruption*, etc.). Those keywords can be extracted from a knowledge base of issues, with their respective related stances and associated terms. This knowledge base should be manually constructed to account for the social context of the population under study, as well as the contingency surrounding political discussion.

We build *user documents*, defined as the concatenation of tweets from each user. We represent each user document u as a vector

$$\vec{u} = [w_0, w_1, \ldots, w_n],$$

where w_i represents the vocabulary word i weighted using TF-IDF [3]:

$$w_i = freq(w_i, u) \times \log_2 \frac{|U|}{|u \in U : w_i \in u|},$$

where U is the set of users, and the vocabulary contains all prototypical keywords as well as all other words used by them. Note that the user document can be built with all tweets and retweets for each user, as well as a subset of both. In particular, we consider tweets and retweets, but not replies to other users, as they are less likely to contribute information to the document. Likewise, for each issue stance we build a stance vector $\vec{s}$, defined as the vectorized representation of tweets containing its prototypical keywords:

$$\vec{s} = [w_0, w_1, \ldots, w_n],$$

with w_i weighted according to TF-IDF with respect to the corpus of user documents.

Using these definitions we can estimate how similar is the language employed by a specific user with the known stances of a specific issue. Formally, we define a user stance with respect to a given sensitive issue as the feature vector $\vec{u}_s$ containing the similarity of user $\vec{u}$ with each issue stance. In this way, we consolidate all similarities in a *user stance vector*:

$$\vec{u}_s = [f_0, f_1, \ldots, f_{|S|}],$$

where S is the set of stances for the all sensitive issues under consideration, and f_i is the cosine similarity between $\vec{u}$ and the issue stance $\vec{s}_i$:

$$\text{cosine\_similarity}(\vec{u}, \vec{s}_i) = \frac{\vec{u} \cdot \vec{s}_i}{\| \vec{u} \| \| \vec{s}_i \|}.$$

Having this representation of user stances, we define the *view gap* with respect to a sensitive issue between two users as the distance between their respective user stance vectors.

Topic Graph. To build the topic graph, we rely on *Latent Dirichlet Allocation*. LDA is a generative topic model that clusters words based on their co-occurrences in documents, and defines latent topics that contribute words to documents. In the past, this model has given reliable results when applied to user documents. Thus, by using LDA we are able to estimate $P(t \mid u)$, for a given latent topic t and a given user document u from the set of users U. The topic graph is an undirected graph $G = \{T, V\}$, where the node set $T = \{t_0, t_1, \ldots, t_k\}$ is comprised of the k latent topics obtained from the application of LDA to the user documents, in the same way as Ramage *et al.* [23]. The edge set is defined as $V = \{v_{i,j} : P(t_i \mid u) \geq \epsilon \wedge \ P(t_j \mid u) \geq \epsilon \ \exists \ u \in U\}$, *i. e.*, two nodes are connected if both corresponding topics contribute (with a minimum probability

of ϵ) to the same user document. Note that edges are weighted according to the fraction of user documents that contributed to it.

Intermediary Topics. To estimate which topics are suitable to be used for recommendation of people of opposing views, we estimate the centrality of each node in the topic graph. In contrast to a previous definition of intermediary topics [16], instead of *betweenness centrality* we compute *current flow closeness centrality* [8] of nodes, which is equivalent to *information centrality* [25]. If the topic graph is considered as an electrical network, with edges replaced with resistances, information centrality is equivalent to the inverse of the average of correlation distances between all possible paths between two nodes. Using this analogy, we expect to measure the degree in which a topic might represent a shared non-challenging interest (*i. e.*, those with the least resistance) between two users. Hence, we redefine *intermediary topics* as topics whose centrality is higher than the median centrality of the entire graph.

In the next section, we evaluate this methodology through a case study of political discussion in Chile.

4. CASE STUDY: ABORTION IN CHILE

In this section we describe a case study where we analyze the issue of abortion in Chile using our methodology. In the context of ongoing campaigns for presidential elections, we crawled tweets from July 24th, 2013 to August 29th, 2013 using the *Twitter Streaming API*. Although we crawled tweets about general political discussion, we did focus crawling and analysis on *abortion*. After the analysis, we statistically evaluate intermediary topics to find how they differ in comparison to non-intermediary topics.

Why Abortion in Chile? The Duality in Discussion. The history of abortion in Chile is long, being declared legal in 1931 and illegal again in 1989. As of 2015, abortion is still illegal, making Chile one of the countries with most severe abortion laws in the world [24]. Abortion in Chile as a sensitive issue has good properties for analysis, as it is constantly being discussed in the political active population. On one hand, 61% of population was estimated to be catholic, and 21% professed another religion, while only 19% of the population was atheist or agnostic [22]. On the other hand, 63% of the Chilean population was in favor of legalization of abortion in 2013 [10]. The occurrence of several protests around public education, same-sex marriage and abortion, among other sensitive issues, are encouraging the usage of micro-blogging platforms and social networks to spread ideas and generate debates (for a discussion on the student movement in Chile see Barahona *et al.* [4]). There is a duality in how the country approach political issues. On one hand, a majority of the population is estimated to have conservative views. On the other hand, a majority of population is in favor of legalization of abortion. Because a growing portion of the population is asking for reforms using social media as a primary communication and organization device, Chile is an ideal scenario for our analysis.

4.1 Dataset Description

Query Keywords and Filtering. Initially, we used *query keywords* about known sensitive issues and hashtags: *abortion* (issue), *education* (issue), *same-sex marriage* (issue), *Sebastián Piñera* (president in 2013), *Michelle Bachelet* (candidate), *Evelyn Matthei* (candidate), among others. When crawling tweets we considered keywords about general political discussion and other sensitive issues (in addition to abortion) because we will consider the relationship between language usage and user stances. We also added emergent hashtags related to news events that happened during the crawling period. For instance, *#yoabortoel25* is about a protest held on July 25th [9]. Figure 1 shows the most frequent terms found in our collection.

Figure 1: Wordcloud of frequent terms in the collection. Green terms were used as query keywords for crawling. Font size is proportional to frequency.

The most prominent words are last names of candidates, namely *Evelyn Matthei*, *Michelle Bachelet*, *Pablo Longueira* and *Laurence Golborne*. The last name of the dictator *Augusto Pinochet* is also prominent. Other prominent keywords are *carabineros* (the police), *censo* (the national level census conducted in 2012, with multiple flaws discovered in 2013), *Transantiago* (public transport system in Santiago), *isapres* (the private health system) and *AFP* (the name of the Chilean private pension system, composed of several *Administrators of Public Founds*). We filtered tweets in other languages than Spanish, tweets that were not geolocated to Chile according to users' self-reported location, as well as tweets about unrelated themes.

Dataset Size. In total, we analyzed 367,512 tweets about political discussion from 57,566 accounts that were geolocated in Chile using a gazetteer. Of those tweets, 18,148 are related to abortion, as they contain at least one prototypical keyword (see Table 1 for the list of keywords related to abortion). The vocabulary size is 38,827, filtering out all keywords that appear in less than 5 tweets.

Pro-Choice and Pro-Life Stances. We manually built a list of words, accounts, and hashtags related to abortion and its two stances. We iteratively explored the dataset to find co-occurrences of prototypical keywords like *abortion*, *#abortolibre* (*free abortion*) and *#noalaborto* (*no to abortion*). For pro-choice and pro-life keywords, the number of seed users and their number of tweets are displayed. These seeds represent whether a user document contained keywords from one stance but not from the other, *e. g.*, a user document that contains at least one pro-choice keyword and no pro-life keywords is considered a pro-choice seed user. As observed in Table 1, the number of pro-choice seed users outnumbers those of pro-life stance (1,934 pro-choice against 338 pro-life). This does not necessarily indicate the proportion of users from both stances. For instance, after performing a manual exploration, some pro-life users who identify themselves as pro-life in their biographies, tend to inject content into pro-choice timelines by publishing tweets with prototypical hashtags from the opposite stance [12].

To build the stance vectors of pro-choice and pro-life stances, we concatenated the tweets of the corresponding seed users of each stance. Then, according to our methodology, we estimated the user stances on abortion by computing the cosine similarity between user vectors and the stance vectors. These similarities are displayed with hexagonal binning in Figure 2, where the x axis represents similarity with the pro-choice stance vector $\vec{s_c}$; and the y axis represents similarity with the pro-life stance vector $\vec{s_\ell}$. We display two charts: one for users who have tweeted about abortion (8,794) on the left, and one that considers all users on the dataset (57,566) on the right. This

Table 1: Keywords used to characterize the pro-choice and pro-life stances on abortion. General keywords plus stance keywords were used to find people who talked about abortion in Twitter. Seeds are users who published tweets with keywords from only one abortion stance.

Stance	Tweets	Seeds	Keywords
Pro-choice	95,173	1,934	#abortolibre, #yoabortoel25, #abortolegal, #yoaborto, #abortoterapeutico, #proaborto, #abortolibresegurogratuito, #despenalizaciondelaborto, #abortoetico, #abortolegal, #abortosinapellido, #derechoadecidir
Pro-life	10,040	338	#provida, #profamilia, #abortoesviolencia, #noalaborto, #prolife, #sialavida, #dejalolatir, #siempreporlavida, provida, #nuncaaceptaremoselaborto, #chilenoquiereabortos, #conabortonohayvoto, #yoasesinoel25, #somosprovida
General Words	–	–	aborto(s), abortista(s), abortados(as), abortivo(a)... ...(tenses of *to abort* in spanish)
Related Hashtags	–	–	#marchaabortolegal, #bonoaborto, #cifrasaborto, #feminismo
Relevant Accounts	–	–	@elardkoch, @siemprexlavida, @quieronacer, @mileschile, @melisainstitute, @ObservatorioGE
Contingency Words	–	–	terapéutico, violada, violación, violaciones, interrupción, inviabilidad, embarazo, embarazada, feto, embrión, fecundación, antiaborto, feminismo

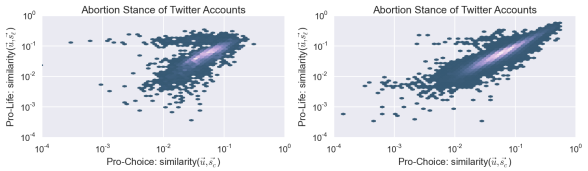

Figure 2: Distributions of user stances based on similarity between user vectors and stance vectors (pro-life and pro-choice). Left: stances of users who tweeted about abortion. Right: stances of all users in the dataset. Both charts use a log-log scale.

is possible because the user stance vectors are constructed using all the vocabulary employed by seed users; hence, they contain valid weights for words unrelated to abortion, but related to additional issues that those users discussed. Under the assumption that sensitive issues have a degree of correlation among stances in different issues, this allows us to estimate a tendency for all users. We define *stance tendency* as:

$$\text{tendency} = \text{cosine\_similarity}(\vec{u}, \vec{s_c}) - \text{cosine\_similarity}(\vec{u}, \vec{s_\ell}).$$

We classify users with tendency ≥ 0 as pro-choice, and pro-life otherwise. The median stance tendency is 0.02, showing a slight tendency towards the pro-choice stance: 54.98% of users are classified as pro-choice, while 45.02% of users are classified as pro-life. Pro-choice users published 10.24 tweets in average, while pro-life users published 10.48 tweets in average.

According to the *Center of Public Studies* [10], 63% of the Chilean population was in favor of legalization of abortion in 2013. Our predicted proportion of user stances does not differ from expectations according to a chi-square test ($\chi^2 = 2.76$, p $= 0.10$). While the Twitter population is not demographically representative of the population, this result indicates that abortion stances are reflected on the micro-blogging platform Twitter.

4.2 Homophily in Two-Way Interactions

Having predicted a stance for each user in the dataset, we are able to evaluate if the interactions in the dataset are homophilic, *i. e.*, we test if users tend to interact with people of the same abortion stance. To do so, we study 2-way interactions. Mentions and retweets are 1-way interactions, where the target user is not necessarily a participant of the interaction. When the target user replies to the mention or the retweet, we consider it a 2-way, bidirectional interaction. To measure homophily, we estimate the aggregated interactions between users in both stances, and compare their inter-stance proportions with the proportions of predicted stances for all

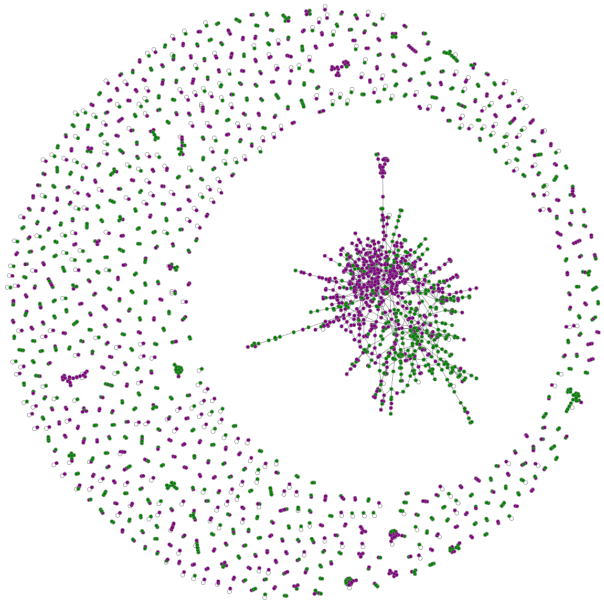

Figure 3: A spring-based graph visualization of two-way user interactions in abortion discussion, where nodes are users. Color encodes abortion stance (purple: pro-choice; green: pro-life).

accounts. If the interaction behavior is unbiased, then the proportion of interactions between stances should not differ from the proportion of users in each stance.

To avoid bias in the estimation, we only considered each pair (u_1, u_2) once per inter-stance interactions. The number of 2-way interactions found for each stance is: pro-choice, 2,234; pro-life, 2,042. The structure of those interactions is visualized in Figure 3, where it can be observed that the largest component has two identifiable clusters, and that small components are prominently of one stance only. The proportions of interactions with the same stance is similar (pro-choice: 76.45%; pro-life: 74.24%). Given the distribution of user stances, in an unbiased population we would expect that each stance would have bidirectional interactions distributed according to the population, *e. g.*, 54.98% of pro-choice users' interactions would be with those of the same stance. A chi-square test indicates that both proportions differ significantly from the expectations (pro-life: $\chi^2 = 29.55$, p < 0.001, Cohen's $w = 0.33$; pro-choice: $\chi^2 = 22.91$, p < 0.001, Cohen's $w = 0.31$), confirming homophilic behavior in the studied population.

4.3 Intermediary Topics

Of all Chileans who published tweets in the case study, we selected a group of 4,077 candidates for analysis of intermediary topics. We considered users that were likely to be *regular users*, *i. e.*, those who follow less than 2,000 accounts and are followed by less than 2,000 (a limit defined by Twitter). This filtering was made because regular people are arguably more prone to discuss their own interests, unlike popular accounts which may be from media outlets, blogs, or celebrities. From those users, we crawled 1,400,582 tweets from December 6th, 2013 until January 3rd, 2014. Jointly with our abortion stance estimation of those users, this makes this dataset useful to test the political diversity of intermediary topics.

We ran LDA with $k = 200$ (a value used before in similar contexts [23]), built the topic graph and estimated information centrality as defined by our methodology. After removing junk topics, which do not contribute to any user document, the graph contains 198 nodes and 6,906 edges. The median centrality is 1.23×10^{-4}, and its maximum value is 1.64×10^{-4}.

We analyze three variables and their relation with centrality, as well as their differences between intermediary and non-intermediary topics: the percent of users that each topic contributes to (Figure 4 Left); the probability of abortion keywords to contribute to each topic (Figure 4 Right), estimated using the LDA model; and the stance diversity (Figure 4 Center), which is the *Shannon entropy* [18] with respect to the predicted abortion stances for all users related to a topic:

$$\text{diversity} = \frac{-\sum_{i=1}^{|S|} p_i \ln p_i}{\ln |S|},$$

where S is the set of stances, and p_i is the probability of stance i, estimated from the fraction of users assigned to each stance according to our methodology.

Proportion of Users. Central topics have much more users than non-central ones: as the number of users increment, centrality does. This is confirmed by a Spearman ρ rank-correlation of 0.99 ($p < 0.001$) between proportion of users and centrality. The maximum proportion of users a topic contributes to is 78.78%, the median value is 0.56% and the mean is 4.13%. The mean for intermediary topics is 7.99%, and for non-intermediary topics 0.26%. This difference is significant according to a Mann-Whitney U test ($U = 12.10$, $p < 0.001$). Hence, intermediary topics are more populated than non-intermediary topics. This is an expected result, because topic graph construction is based on how topics are related to users.

Stance Diversity. Nodes with high stance diversity can have low centrality, but they concentrate in the upper middle of the chart. The maximum diversity of a topic is 1, its median value is 0.97 and its mean is 0.91. The mean for intermediary topics is 0.96, and for non-intermediary topics 0.86. This difference is significant according to a Mann-Whitney U test ($U = 3.30$, $p < 0.001$), meaning that intermediary topics are more likely to contain a greater diversity of people with different views on abortion than non-intermediary topics.

Topical Probability of Abortion-Related Vocabulary. Using our set of prototypical keywords, we can estimate the probability of abortion-related vocabulary to contribute to specific topics $P(A \mid t)$, where A is the set of keywords, and t is the target topic:

$$P(A \mid t) = \sum_{i=1}^{|A|} P(w_i \mid t),$$

where w_i is the ith word in A. Note that the LDA model allows us to estimate $P(w_i \mid t)$ directly. Figure 5 displays the distributions and

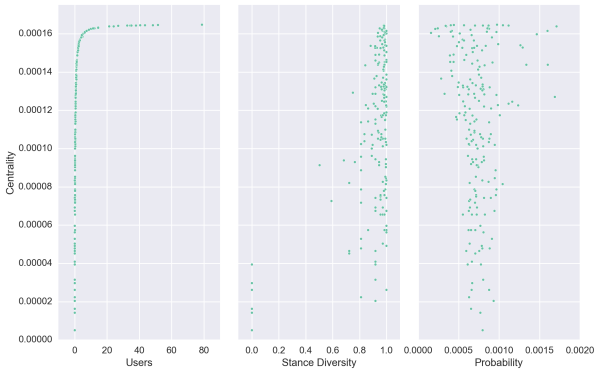

Figure 4: Relationship between topic information centrality [8] and the percent of users the topic contributes to (left), the abortion-stance diversity estimated with *Shannon entropy* [18] (center), and the probability of abortion-related keywords to contribute to each topic (right).

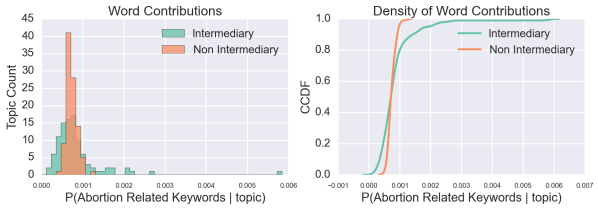

Figure 5: Left: Histograms of abortion-related keywords contributions to intermediary and non-intermediary topics. Right: Cumulative Density Function .

Complimentary Cumulative Density Functions (CCDFs) of probabilities for intermediary and non-intermediary topics. Although the distribution chart hints a potential difference, this difference is not significant according to a Mann-Whitney U test ($U = -0.59$, $p = 0.55$).

5. DISCUSSION

In this paper we have confirmed that intermediary topics do exist and are measurable. We have improved the definition of intermediary topics by Graells-Garrido *et al.* [16], as well as quantified homophilic discussion and the differences between intermediary and non-intermediary topics. In particular, we have found that intermediary topics are more likely to contain a diverse set of users in terms of political stances, and thus, are suitable for use in recommendation of people of opposing views. We devise these topics as important features that could help to avoid *cognitive dissonance* [14] in users when facing recommendations. Although our results apply to the studied community from Chile, the methods used are generalizable to other communities as long as there are known prototypical keywords for the sensitive issues to be studied.

In addition, the way in which we quantified homophily can be used as a metric to evaluate the polarization in discussion around specific political issues. In our case study, polarization of stances had considerable effect sizes (measured with Cohen's w), meaning that discussion in Chile around abortion is highly polarized, a result supported by national surveys of political discussion [22, 10].

A question that arises regarding intermediary topics is: does the definition of intermediary topics hold when considering general

political views instead of a specific sensitive issue? We propose that it does because by definition intermediary topics only rely on the estimation of information centrality [8]. However, this is left for future work. Additionally, future work will consider the incorporation of intermediary topics into a recommender system to be evaluated with users, as well as the interaction of intermediary topics with social- and content-based signals.

Acknowledgments. We thank the anonymous reviewers for their helpful feedback. This work was partially funded by Grant TIN2012--38741 (Understanding Social Media: An Integrated Data Mining Approach) of the Ministry of Economy and Competitiveness of Spain.

References

[1] Jisun An, Daniele Quercia, and Jon Crowcroft. "Why individuals seek diverse opinions (or why they don't)". In: *Proceedings of ACM Web Science*. 2013, pp. 11–15.

[2] Solomon E Asch. "Forming impressions of personality." In: *The Journal of Abnormal and Social Psychology* 41.3 (1946), p. 258.

[3] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information retrieval: the concepts and technology behind search, 2nd. Edition*. Addison-Wesley, Pearson, 2011.

[4] Matías Barahona, Cristóbal García, Peter Gloor, and Pedro Parraguez Ruiz. "Tracking the 2011 student-led movement in Chile through social media use". In: *Collective Intelligence 2012* (2012).

[5] Pablo Barberá. "Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data". In: *Political Analysis* 23.1 (2015), pp. 76–91.

[6] David M Blei, Andrew Y Ng, and Michael I Jordan. "Latent Dirichlet Allocation". In: *The Journal of Machine Learning Research* 3 (2003), pp. 993–1022.

[7] Antoine Boutet, Hyoungshick Kim, and Eiko Yoneki. "What's in Twitter, I know what parties are popular and who you are supporting now!" In: *Social Network Analysis and Mining* 3.4 (2013), pp. 1379–1391.

[8] Ulrik Brandes and Daniel Fleischer. "Centrality measures based on current flow". In: *STACS 2005* (2005), pp. 533–544.

[9] María Jesús Ibáñez Canelo. "El control de los cuerpos de las mujeres es algo medular en la política patriarcal capitalista: entrevista a Soledad Rojas, feminista chilena". In: *Comunicación y Medios* 30 (2015). [In Spanish. Title translation: The control of women's bodies is something core in capitalist patriarchal politics: interview with Soledad Rojas, Chilean feminist.]

[10] CEP. *National Survey of Public Opinion, September–Octuber 2013*. http://www.cepchile.cl/1_5388/doc/estudio_nacional_de_opinion_publica_septiembre-octubre_2013.html. [Online; accessed April 2015]. 2013.

[11] Jilin Chen, Werner Geyer, Casey Dugan, Michael Muller, and Ido Guy. "Make new friends, but keep the old: recommending people on social networking sites". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM. 2009, pp. 201–210.

[12] Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. "Political polarization on twitter." In: *International Conference on Weblogs and Social Media*. 2011.

[13] Michael D Conover, Bruno Gonçalves, Jacob Ratkiewicz, Alessandro Flammini, and Filippo Menczer. "Predicting the political alignment of twitter users". In: *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom)*. IEEE. 2011, pp. 192–199.

[14] Leon Festinger. *A theory of Cognitive Dissonance*. Vol. 2. Stanford University Press, 1962.

[15] Eduardo Graells-Garrido and Mounia Lalmas. "Balancing Diversity to Counter-measure Geographical Centralization in Microblogging Platforms *(short paper)*". In: *25th ACM Conference on Hypertext and Social Media* (2014).

[16] Eduardo Graells-Garrido, Mounia Lalmas, and Daniele Quercia. "People of opposing views can share common interests". In: *Proceedings of the companion publication of the 23rd international conference on World Wide Web (poster)*. International World Wide Web Conferences Steering Committee. 2014, pp. 281–282.

[17] John Hannon, Mike Bennett, and Barry Smyth. "Recommending twitter users to follow using content and collaborative filtering approaches". In: *Proceedings of the fourth ACM Conference on Recommender Systems*. ACM. 2010, pp. 199–206.

[18] Lou Jost. "Entropy and diversity". In: *Oikos* 113.2 (2006), pp. 363–375.

[19] Miller McPherson, Lynn Smith-Lovin, and James M Cook. "Birds of a feather: Homophily in social networks". In: *Annual review of sociology* (2001), pp. 415–444.

[20] Eli Pariser. *The filter bubble: What the Internet is hiding from you*. Penguin UK, 2011.

[21] Marco Pennacchiotti and Ana-Maria Popescu. "Democrats, republicans and starbucks afficionados: user classification in twitter". In: *Proceedings of the 17th ACM SIGKDD international conference on Knowledge Discovery and Data Mining*. ACM. 2011, pp. 430–438.

[22] Pontificia Universidad Católica de Chile. *Encuesta Bicentenario UC-Adimark, 2014*. http://encuestabicentenario.uc.cl/. [In Spanish; Online; accessed April 2015]. 2014.

[23] Daniel Ramage, Susan Dumais, and Dan Liebling. "Characterizing microblogs with topic models". In: *International Conference on Weblogs and Social Media*. Vol. 5. 4. 2010, pp. 130–137.

[24] Bonnie L Shepard and Lidia Casas Becerra. "Abortion policies and practices in Chile: ambiguities and dilemmas". In: *Reproductive Health Matters* 15.30 (2007), pp. 202–210.

[25] Karen Stephenson and Marvin Zelen. "Rethinking centrality: Methods and examples". In: *Social Networks* 11.1 (1989), pp. 1–37.

[26] Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. "Twitterrank: finding topic-sensitive influential twitterers". In: *Proceedings of the third ACM WSDM*. 2010, pp. 261–270.

[27] Sarita Yardi and Danah Boyd. "Dynamic debates: An analysis of group polarization over time on twitter". In: *Bulletin of Science, Technology & Society* 30.5 (2010), pp. 316–327.

Analysis of Sentiment Communities in Online Networks

Davide Feltoni Gurini, Fabio Gasparetti,
Alessandro Micarelli and Giuseppe Sansonetti
Roma Tre University
Department of Engineering
Via della Vasca Navale 79
Rome, 00146 Italy
{feltoni,gaspare,micarel,gsansone}@dia.uniroma3.it

ABSTRACT

This article reports our experience in developing a recommender system (RS) able to suggest relevant people to the target user. Such a RS relies on a user profile represented as a set of weighted concepts related to the user's interests. The weighting function, we named *sentiment-volume-objectivity (SVO) function*, takes into account not only the user's sentiment toward his/her interests, but also the volume and objectivity of related contents. A clustering technique based on modularity optimization enables us to identify the latent sentiment communities. A preliminary experimental evaluation on real-world datasets from Twitter shows the benefits of the proposed approach and allows us to make some considerations about the detected communities.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: [Information Filtering]

General Terms

Algorithms, Experimentation

1. BACKGROUND

The rationale behind this work is that users may share similar interests, but have different opinions about them. Therefore, we extend the traditional approaches to user recommendation through sentiments and opinions extracted from the user-generated content in order to improve the accuracy of suggestions. This way, we can identify latent *sentiment communities* of users. As far as we know, there are few works on considering users' attitudes for community detection or user recommendation. In [6] the authors formulate the problem of sentiment community discovery as a semidefinite programming (SDP) problem and solve it through an SDP-based rounding method. Nguyen *et al.* [5] address the problem of clustering blog communities into groups based on users' sentiments, and propose a non-parametric clustering algorithm for its solution. Yang and Manandhar [7] propose two community discovery models by combining social links, author based topics and sentiment information to detect communities with different sentiment-topic distributions. Unlike previous works, we consider not only the target user's attitudes toward his/her interests, but also the volume and objectivity of related generated contents.

2. PROPOSED APPROACH

Traditional approaches to user recommendation rely on the definition of a similarity measure between two users u and v. Given the target user u, the ranked list of suggested users corresponds to the set of users v that maximize the aforementioned measure. Content-based approaches on Twitter [1] define this measure by analyzing user tweets. Concepts dealt with by a user are identified through *hashtags* contained in his/her tweets, namely, the metadata tags that are used in Twitter to indicate the context or the flow a tweet is associated with. Thus, we define the profile p of the user u as the set of weighted concepts:

$$p(u) = \{(c, \omega(u,c)) | c \in C_u\} \quad (1)$$

where $\omega(u,c)$ is the relevance of the concept c for the user u, and C_u is the set of concepts cited by the user u. The user profile representation is generated by monitoring the user activity, that is, all the tweets included in the observation period. Afterwards, given two users u and u, and their profiles $p(u)$ and $p(v)$, the similarity function is defined in terms of cosine similarity:

$$sim(u,v) = sim(p(u),p(v)) = \frac{\sum_{c \in C_u \cup C_v} \omega(u,c) \cdot \omega(v,c)}{\sqrt{\sum_{c \in C_u} \omega(u,c)^2} \cdot \sqrt{\sum_{c \in C_v} \omega(v,c)^2}} \quad (2)$$

where C_u and C_v are the concepts in the profiles of users u and v, respectively. The idea behind this work is that taking into account users' attitudes towards their interests can yield benefits in recommending friends to follow. Specifically, we consider three contributions: 1) $S(u,c)$, that is, the sentiment expressed by the user u for the concept c; 2) $V(u,c)$, that is, how much he/she is interested in that concept; 3) $O(u,c)$, that is, how much he/she expresses objective comments on it. The details regarding the computation of such contributions can be found in [2]. Based on those contributions, we propose a weighting function, we called *sentiment-volume-objectivity (SVO)* function, that takes into account

[1] `twitter.com`

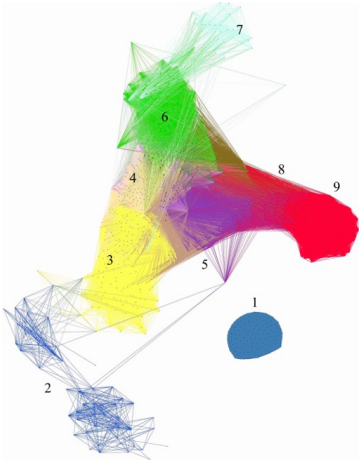

Figure 1: Communities for the *Apple* concept, detected through the SVO profiling and clustering.

all of them. It is defined as follows:

$$SVO(u,c) = \alpha S(u,c) + \beta V(u,c) + \gamma O(u,c) \qquad (3)$$

where α, β, and γ are three constants $\in [0,1]$, such that $\alpha + \beta + \gamma = 1$. The function $SVO(u,c) \in [0,1]$ is the weighting function $\omega(u,c)$ that appears in Equations 1 and 2. Once the similarities between users are computed, we build a graph for each concept as follows: if the similarity value between users exceeds a threshold value θ, we consider an edge between them. The optimal value for θ was determined through a gradient descent algorithm that maximizes the recommender precision. Such value was 0.8. Afterwards, a clustering algorithm based on modularity optimization [1] allows us to detect the latent communities for the considered concept c.

3. EXPERIMENTAL EVALUATION

Experimental tests were performed on three datasets gathered from Twitter through its APIs [2] by searching for specific hashtags. $Dataset_1$ was obtained during the 2013 Italian political elections. We retrieved the Twitter streams about politician leaders and Italian parties from January 25th to February 27th. The final dataset counted 1,085,121 tweets in Italian language and 70,977 unique users. $Dataset_2$ was obtained searching for hashtags and keywords representing the most important mobile tech companies such as Samsung, Apple, and Nokia. The dataset was gathered from September 2014 to February 2015 considering only Italian tweets, and counted 3,511,455 tweets from 181,000 users. $Dataset_3$ was obtained analyzing English tweets on the automotive landscape. To this aim, we searched terms such as Audi, BMW, and Ferrari. The collection set, gathered from December 2014 to February 2015, counted 2,915,131 tweets from 110,350 users. Figure 1 shows the different communities detected for the Apple concept ($dataset_2$). The bottom, right, isolated, community marked with number one includes users not interested in Apple ((i.e., their SVO value is zero). The bottom, left, community with number two consists of users with low interest (i.e., low value of volume) and negative sentiment. The community labeled with number three is characterized by high interest and negative sentiment. The central communities identified with numbers four, five, six, and seven encompass users with high interest and positive sentiment (with different SVO combinations for each community). Users belonging to communities eight and nine have high values of objectivity, namely, they generate objective contents with few opinions. Among such users, for example, we find online newspapers such as BBC and CNN.

[2] `dev.twitter.com`

Table 1: A comparison among different techniques

User Recommender	$Dataset_1$	$Dataset_2$	$Dataset_3$
Our Approach	**0.177**	**0.195**	**0.185**
S1-Twittomender	0.130	0.118	0.115
VSM (Hashtag)	0.127	0.099	0.105

Table 2: SVO parameters for the three datasets

Parameter	$Dataset_1$	$Dataset_2$	$Dataset_3$
α	0.45	0.25	0.28
β	0.45	0.50	0.52
γ	0.10	0.25	0.20

In order to evaluate our approach, we relied on the *homophily* [4] phenomenon, that is, the tendency of individuals with similar characteristics to associate with each other. For each dataset, we selected 1000 users that (i) posted at least 50 tweets in the observed period, and (ii) had more than 30 friends and followers. Table 1 reports the results in terms of *Success at Rank 10 (S@10)* of a comparative analysis of our system with two state-of-the-art functions: 1) a content-based function, called *S1-Twittomender* [3], where users are profiled through the content of their tweets; and 2) a *VSM (Hashtag)* function representing cosine similarity in a vector space model, where vectors are weighted hashtags. Table 2 shows the values of SVO parameters α, β, and γ that maximize the performance of the recommender. Such values were determined through a mini-batch gradient descent algorithm. Based on the proposed model and the used datasets, these weights appear to highlight the contribution of volume and sentiment in $dataset_1$, and objectivity in $dataset_2$ and $dataset_3$. This can be explained because $dataset_2$ (technology) and $dataset_3$ (automotive) are likely to contain more news and articles with few opinions and sentiments than $dataset_1$ (politics).

4. CONCLUSIONS AND FUTURE WORK

In this article, we have presented some results of our research work whose aim is to exploit the implicit sentiment analysis for improving the performance of user recommenders. In particular, we have reported some preliminary considerations on the sentiment communities that our approach allows us to identify.

Our work is still at an exploratory stage, so several its aspects have to be further developed. Among others, we plan to perform an in-depth sensitivity analysis to study how user preferences, social interactions, and dataset characteristics can affect SVO parameters α, β, and γ. Their role is indeed crucial, since they define the contribution extent of sentiment, volume, and objectivity that determine the distribution of users in sentiment communities.

5. REFERENCES

[1] V. Blondel, J. Guillaume, R. Lambiotte, and E. Mech. Fast unfolding of communities in large networks. *J. Stat. Mech*, 2008.

[2] D. F. Gurini, F. Gasparetti, A. Micarelli, and G. Sansonetti. A sentiment-based approach to twitter user recommendation. In *Proc. of the Fifth RSWeb Workshop co-located with the Seventh ACM RecSys Conference, Hong Kong, China*, 2013.

[3] J. Hannon, M. Bennett, and B. Smyth. Recommending twitter users to follow using content and collaborative filtering approaches. *Proceedings of the 4th ACM RecSys Conference*, 26-30(10):8, 2010.

[4] M. McPherson, L. S. Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1):415–444, 2001.

[5] T. Nguyen, D. Q. Phung, B. Adams, and S. Venkatesh. A sentiment-aware approach to community formation in social media. In *ICWSM*. The AAAI Press, 2012.

[6] K. Xu, J. Li, and S. S. Liao. Sentiment community detection in social networks. In *Proc. of the 2011 iConference*, pages 804–805, New York, NY, USA, 2011.

[7] B. Yang and S. Manandhar. Community discovery using social links and author-based sentiment topics. In *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on*, pages 580–587, 2014.

Retrieving Relevant Conversations for Q&A on Twitter

Jose Miguel Herrera
PRISMA Research Group
Department of Computer Science
University of Chile, Chile
jherrera@dcc.uchile.cl

Denis Parra
Department of Computer Science
Pontificia Universidad Católica de Chile
dparra@ing.puc.cl

Barbara Poblete
PRISMA Research Group
Department of Computer Science
University of Chile
bpoblete@dcc.uchile.cl

ABSTRACT

Community Question and Answering (Q&A) sites provide special features for asking questions and receiving answers from users on the Web. Nevertheless, Web users do not restrict themselves to posting their questions exclusively in these platforms. With the massification of on-line social networks (OSN) such as Twitter, users are increasingly sharing their information needs on these websites. Their motivation for doing so is to obtain a timely and reliable answer from their personal community of trusted contacts. Therefore, daily on Twitter, there are hundreds of thousands of questions being shared among users from all over the world. Many of these questions go unanswered, but also an important number receive relevant and complete replies from the network. The problem is that due to the volatile nature of the streaming data in OSN and the high arrival rates of messages, valuable knowledge shared in this Q&A interaction lives very shortly in time. This produces high redundancy and similarity in questions which occurs consistently over time. Following this motivation we study Q&A conversations on Twitter, with the goal of finding the most relevant conversations posted in the past that answer new information needs posted by users. To achieve this we create a collection of Q&A conversation threads and analyze their relevance for a query, based on their contents and relevance feedback from users. In this article, we present our work in progress which includes a methodology for retrieving and ranking Q&A conversation threads for a given query. Our preliminary findings show that we are able to use historical conversation on Twitter to answer new queries posted by users. We observe that in general the asker's feedback is a good indicator of thread relevance. Also, not all of the feedback features provided by Twitter are equally useful for ranking Q&A thread relevance. Our current work focuses on determining empirically the best ranking strategy for the recommendation of relevant threads for a new user question. In the future we seek to create an automatic Q&A knowledge base that is updated in real-time that allows for preserving and searching human understanding.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Relevance feedback; H.1.2 [**User/Machine Systems**]: Human factors

General Terms

Experimentation, Human Factors, Algorithms.

Keywords

Twitter, Recommendation, Ranking, Q&A, Threads.

1. INTRODUCTION

Question and Answering (Q&A) websites allow users to ask questions and receive answers from a diverse group of people. These kinds of sites accumulate knowledge providing a valuable resource of information that cannot be easily obtained using Web search engines. Popular Q&A sites are *Yahoo! Answers*[1] and *Stack Overflow*[2], which are specially designed to generate Q&A interaction among users. One property of these kinds of sites is the ability to choose the *best answer* for a particular question through a community-wise voting system. In general, the *best answer* selection is based on the amount of positive votes for a user's answer. This feature is one of the most important characteristics in these kinds of sites, because these highly voted answers will solve similar future questions. With this social mechanism that allows for collecting good-quality questions and answers, Q&A sites motivate people to come back for obtaining almost immediate answers to their information needs.

Although Q&A sites are popular, users also ask a significant volume of questions online in other non-specialized but more popular networking platforms, such as Facebook or Twitter. This behavior might seem counterintuitive, especially on Twitter, due to the volatile nature of its information stream and the lack of special incentives for motivating users' answers available in Q&A sites – such as badges and enhanced rights for active users. This seemingly suboptimal behavior might be explained by the observations of Morris et al. [7], who showed that the main motivation that users have for asking questions online on Q&A platforms is to receive quick and trustworthy answers - something that can be potentially achieved in a massive microblogging site like Twitter. Considering this background, we conducted a preliminary analysis which indicated that around 10% of the Twitter stream corresponds to Q&A messages (similar measures were obtained by [4, 6]). Moreover, a rough inspection at Q&A conversation threads on Twitter yields high redundancy of questions over time, meaning that there is a high chance of finding answers to newly asked questions. This trend of Q&A usage in Twitter indicates a increasing potential for fulfilling current users' information needs based on similar questions already answered in the past. Previous work shows research

[1]http://answers.yahoo.com: General-purpose Q&A.
[2]http://www.stackoverflow.com: Software-development Q&A.

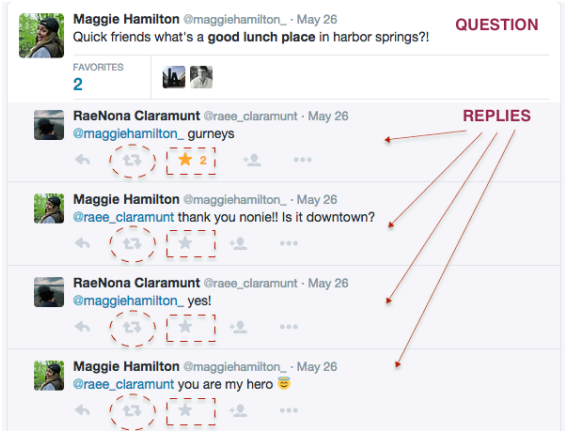

Figure 1: A conversation thread on *Twitter* formed by one question and four answers. Tweets can be *Re-tweeted* (dotted circle) and *Favorited* (dotted square).

on conversations in Twitter, such as [1, 2, 5], but none of them focuses on building a knowledge repository of Q&A in Twitter to answer a question's query. Other researchers have tried to automatically reply to unanswered questions by matching similar questions already answered in the past [8, 10], but they employed the *Yahoo! Answers* platform, and the challenge in Twitter is more complex due to the lack of explicit mechanisms to tell good from bad answers as in Q&A sites. For these reasons, we address the task of creating a method for retrieving the most relevant historical conversation threads which can answer a given query q, by leveraging Twitter as Q&A repository.

We study how the combination of questions, their replies and Twitter social features (retweets, replies, favorites, etc.) can be useful for determining whether a conversation thread triggered by a question is relevant in terms of information quality, to the particular conversation topic. In particular, we aim to investigate the following research questions:

- **RQ1:** how can we determine whether a conversation thread was resolved (answered) on Twitter? In other words, which factors or features are most relevant to determine that? and,
- **RQ2:** can we recommend relevant conversation threads made in the past to answer a new question? Can we build a ranking model with the most relevant features of **RQ1**?

We define the relevance of a conversation thread in terms of its likelihood of providing a complete answer and resolving the information need. Then, to evaluate the importance of each conversation we employed a relevance measure which is based on the feedback provided by the user asking the question on Twitter. Our preliminary findings show that the feedback of the asker plays an important role to evaluate whether a conversation thread had good quality, i.e., whether it was resolved or not. Nevertheless, the noisy nature of tweets makes them complex to analyze, making our problem difficult.

The main contributions of this paper are: **(1)** proposing a methodology for obtaining a set of ranked historical conversation threads

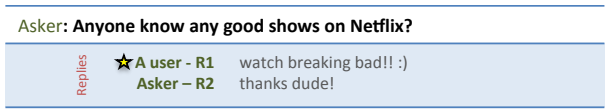

Figure 2: An example of a thread with two replies (R1 and R2) where an *Asker* asks a question, a *User* replies, and finally the *Asker* replies back. The star means the *Asker* has marked that reply as Favorite.

that answer a given question, and **(2)** identifying the main characteristics that influence the quality of a conversation thread. To the best of our knowledge, the method proposed in this article is the first attempt to rank conversation threads based on feedback in Twitter. The applications of this work can be useful for any social network with interaction among users to enhance the results on search. Also, we can use this approach to build a question and answer repository website based on Twitter.

The rest of the paper is structured as follows: Section 2 describes our methodology to identify and rank Q&A threads based on questions asked on Twitter, Section 3 provides details of our preliminary experiments, such as the dataset and uses cases where our model works appropriately, and Section 4 provides a brief summary of our initial expectations and future work.

2. RANKING Q&A THREADS

In this section we present our preliminary methodology for retrieving and ranking relevant Q&A conversation threads for a previously unknown query on Twitter. We define a Q&A conversation thread (hereinafter *threads*), as a conversation on Twitter in which the initial tweet is phrased as a question. We define the relevance of a thread in terms of *how effective the complete conversation is at answering the information need posed in the initial tweet of the conversation*. See Figure 1 for an example of a Q&A conversation thread.

In order to identify features that characterize whether a conversation thread has already resolved an information need, we manually inspected several hundreds of conversation threads. This analysis brings us to consider that replies in conversation threads are important at the moment of establishing the relevance of the conversation. In particular, given Twitter's relevance feedback options, tweets in a conversation thread can be marked by users as *Favorite* and/or *Retweeted*, where the first indicates a special preference and the second indicates that the content has been re-posted by a user. In particular, we observe that the feedback provided by the user who posted the question which initiates a thread, called the *asker*, plays an important role indicating the relevance of the thread. Our intuition is that since the asker is very interested in obtaining a good answer to his/her query, a frequent use of Twitter's relevance feedback features will indicate a higher satisfaction. Figure 2 shows a simple instance where the asker gives feedback in a thread. In this case, with an option to determine the level of satisfaction of the asker with a thread, we can evaluate the second reply "thanks dude!" of the *asker* using Sentiment Analysis (using NLTK tools[3] we obtain the reply has a positive polarity of $67,13\%$). This follows a similar approach by Pelleg et al. [8] for Yahoo! Answers. In the example, if the asker additionally marks the first reply as a *Favorite* (followed by a positive answer) this provides a stronger

[3]Natural Language Toolkit, http://text-processing.com

Figure 3: Case 1. Given a question q^*, we show the top-2 relevant threads. The stars mean that the tweet was marked as Favorite by the Asker.

indication of satisfaction with the reply. We call this type of behavior *positive reinforcement feedback* (PRF), in which the asker indicates its approval for replies to his/her question. In our initial inspection of our dataset we have identified at least 5 other similar types of PRF which are recurrent over time in Q&A threads.[4]

More formally, given a new question q^* we retrieve a set of its top-k similar conversation threads. We do so initially by retrieving threads with the highest cosine similarity of their initial tweet q^*. We denote this set of similar threads as $T = \{th_1, \ldots, th_k\}$. Each thread is given by $th_j =< q_j, R_j >$, in which q_j is the initial tweet or question of th_j and R_j is the set of replies received for q_j. Then, for each thread th_j we compute its absolute relevance $rel(th_j)$ that indicates the level of satisfaction of the asker of q_j with the overall replies received in th_j. Initially we estimate $rel(th_j)$ as:

$$rel(th_j) = count\ of\ positive\ reinforcement\ instances\ in\ th_j$$

Using the value of $rel(th_j)$ we re-order the set T obtained for q^*. Just we take the top-k elements with highest $rel(th_j)$. We do not use a threshold value because each thread presents a different levels of feedback. Hence, we can not define a fixed value.

3. PRELIMINARY EXPERIMENTS

In this section we present the dataset used in our experiments, preliminary results and some findings.

Dataset. In order to show evidence of the usefulness of our proposal, we collected a dataset of tweets in English language. This preliminary Twitter dataset contains 721 questions (q^*) and 152, 721 conversation threads (th_j). We created this collection from the public Twitter API. Since our goal is to have sets of similar questions in the dataset and question threads are very sparse in the public stream, we conducted a focalized crawl for threads. This is, we retrieved questions and similar question threads using the following iterative process: 1) we search in Twitter for a list of common words used in questions, 2) filter all of the results (tweets) that corresponded to a question q^*, and 3) for each question q^*, perform an additional search to retrieve similar questions-threads th_j. The full process (1)-(3) was conducted between March 31, 2015 and April 27, 2015.

The first and second steps were carried out through the **Streaming API**[5] using a traditional rule-based approach [4, 7, 11]. Also, we have defined certain rules of questions that we need, because not just any question is useful in our task. We collect questions that require answers (information needs or factual knowledge), questions that are not affected by time, recommendation questions, suggestion requests, questions in English, and opinion questions. For instance, we keep in our dataset questions such as: "does anyone know cheap places to stay in London?", "does anybody can recommend me good restaurants in Santa Monica?". On the other hand, we discard questions such as: "anyone got an iPhone 5 for sale?", "Anyone know what time the mayweather fight starts??". The first is not a factual knowledge question and the second is affected by time (we have proposed to include these kinds of questions in future work).

Ranking conversation threads. The third step is to build the set of similar past threads th_j of q^*. Since Twitter API restricts obtaining the complete thread, we must first retrieve similar tweets and then the replies, if they exist. We have retrieved all the q_j that are similar to q^* from the **Search API**[6] (the retrieval is based on the main keywords of q^*). Then, we retrieved the replies R_j of q_j to build the thread structure. We have adapted a development made by Adrian Crepaz[7] that can get replies through the Twitter mobile webpage. Finally, we calculate the relevance $rel(th_j)$ for each thread based on the PRF.

3.1 Q&A Ranking Examples

By both automatic analysis and manual inspection of our dataset, we identified common patterns of Q&A conversation threads. In this subsection we describe three of the most recurrent examples and how our ranking methodology works in each case.

Case 1. Figure 3 shows the top 2 similar threads retrieved by our approach sorted by relevance (high to low), given the initial question q^*: "Anyone have any good book recommendations???" (it was taken literally). The first thread contains three replies and two of them were marked as Favorites by the asker (see the starts). Notice that the first reply R1 (of the first thread) was marked as Favorite by the asker and followed by the asker (R2) with positive sentiment. The sentiment analysis of R2 gave us 76% of probability that the text presents positive polarity. That means that the thread presents PRF. The reply R3 of the first thread was marked as favorite by the asker but it is not followed by any tweet of the asker. On the other hand, the second thread just presents a positive sentiment in the reply R6 ("ty" means "thank you"). Although the asker uses feedback elements (positive expression in the reply R6), the second thread does not present the structure to be PRF. Hence, the relevance is lower than the first thread.

Case 2. Figure 4 presents another recurrent case. Given an initial question q^*, the threads retrieved are just the initial tweet q_i, without replies. But if we observe, the retrieved tweets still can answer the question q^*. When this occurs, we sort the tweets depending

[4]We do not enter in details at this moment of all of the types of PRF, given that we are presenting our preliminary findings in brief format.

[5]The streaming API captures 1% of the Twitter volume in real time.
[6]The search API retrieves tweets posted within a week of the time the query was issued.
[7]http://adriancrepaz.com/twitter_conversations_api

Question q^*: **anyone got some good free online games ?**

#	Retrieved Threads
1	**Asker**: Tower defense Inferno, is a good simple tower defense game, have fun :) http://t.co/HkS9CYPayE #fb
	********* NO REPLIES **********
2	**Asker**: http://t.co/wl2vBKptfL what are some good (preferably free) multiplayer games, or games that can be played online with others via lan or...
	********* NO REPLIES **********
3	**Asker**: Utica Comets Game Streams?: Anyone know where I could stream the playoff games for free online? Its good watc... http://t.co/oBM3TdlmXB
	********* NO REPLIES **********
4	**Asker**: [Video & Online Games] Open Question : What is a Good Free Online Fighting Game?: By which I mean something in the vein of Street Fighter
	********* NO REPLIES **********
5	**Asker**: does anyone know some good multiplayer online games that are free
	********* NO REPLIES **********

Figure 4: Case 2. The top-5 threads retrieved are just tweets (without replies), but we can sort them by URLs.

on whether they contain URLs. Chen et al. [3] shows that the tweet relevance is high when it contains a URL. In our dataset, the amount of threads without replies are 69.9%. We notice that after the third tweet the tweets do not clearly reply to the initial question $q*$.

Case 3. The conversation threads could have high relevance if they had more instances of PRF within the same thread. Figure 5 shows this case, where the thread presents PRF twice in one thread. The reply R1 has been marked as Favorite by the Asker. The reply R2 was made by the Asker with 60% of positive sentiment. Hence, the replies R1-R2 present PRF. The replies R3-R4 also present PRF (R3 was marked as Favorite by the Asker and reply R4 returned 54% of positive sentiment). Although reply R5 has a Favorite made by the Asker, the reply R6 has a neutral sentiment. Therefore, they do not present PRF.

4. CONCLUSIONS AND FUTURE WORK

The cases presented in this chapter provide evidence of how our method is used for retrieving and ranking historical conversations threads in order to answer recent questions. This is preliminary work and there is much left to do in the future, such as, validation based on human judgement. The main goals of the evaluation are: (1) whether our positive reinforcement instances are accurate to correctly classify relevant threads (**RQ1**), and (2) whether our ranking approach supports recommendation of relevant threads (**RQ2**).

Such results can lead us to to determining the best model and highlight which Twitter features are more relevant to our task. If we detect that there are several features that can influence in determining the relevance of a thread, we propose use machine learning techniques to automatically construct the ranking model based on the aforementioned features.

Question q^*: **Does anyone know of a website where I can watch movies? :-)**

#	Retrieved Threads	
1	**Asker**: website to watch movies online?	
Replies	★**User - R1**	projectfreetv.so
	Asker - R2	Sound, cheers
	★**User - R3**	putlocker.is
	Asker - R4	sound
	★**User - R5**	Showbox
	Asker - R6	Cheers

Figure 5: Case 3. Two types of *positive reinforcement feedback* (PRF) in one thread: R1 was marked as *Favorite* by asker and R2 has positive sentiment. The same with replies R3-R4. The reply R5 has a favorite but R6 presents a neutral sentiment.

5. REFERENCES

[1] D. Boyd, S. Golder, and G. Lotan. Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. In *System Sciences (HICSS), 2010 43rd Hawaii International Conference on*, 2010.

[2] J. Chen, R. Nairn, and E. H.-h. Chi. Speak little and well: recommending conversations in online social streams. *CHI*, 2011.

[3] J. Chen, R. Nairn, L. Nelson, M. Bernstein, and E. Chi. *Short and tweet: experiments on recommending content from information streams*. ACM, 2010.

[4] M. Efron and M. Winget. Questions are content: a taxonomy of questions in a microblogging environment. In *ASIS&T '10: Proceedings of the 73rd ASIS&T Annual Meeting on Navigating Streams in an Information Ecosystem*, 2010.

[5] C. Honey and S. C. Herring. Beyond Microblogging: Conversation and Collaboration via Twitter. In *System Sciences, 2009. HICSS '09. 42nd Hawaii International Conference on*, 2009.

[6] B. Li, X. Si, M. R. Lyu, I. King, and E. Y. Chang. Question identification on twitter. In *CIKM '11: Proceedings of the 20th ACM international conference on Information and knowledge management*, 2011.

[7] M. R. Morris, J. Teevan, and K. Panovich. What do people ask their social networks, and why?: a survey study of status message q&a behavior. In *CHI '10: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2010.

[8] D. Pelleg, O. Rokhlenko, M. Shovman, I. Szpektor, and E. Agichtein. The Crowd is Not Enough: Improving User Engagement and Satisfaction Through Automatic Quality Evaluation. In *SIGIR '15: Industry Track*, 2015.

[9] E. M. Rodrigues and N. Milic-Frayling. Socializing or knowledge sharing?: characterizing social intent in community question answering. *CIKM*, 2009.

[10] A. Shtok, G. Dror, Y. Maarek, and I. Szpektor. Learning from the past: answering new questions with past answers. In *WWW '12: Proceedings of the 21st international conference on World Wide Web*, 2012.

[11] Z. Zhao and Q. Mei. Questions about Questions: An Empirical Analysis of Information Needs on Twitter. In *WWW*, 2013.

Persona-ization: Searching on Behalf of Others

Paul Bennett
Microsoft Research
One Microsoft Way
Redmond, WA
pauben@microsoft.com

Alexander Fishkov[1]
St. Petersburg Academic University
Academy of Sciences
St. Petersburg, Russia
alexander.fishkov@yandex.ru

Emre Kıcıman
Microsoft Research
One Microsoft Way
Redmond, WA
emrek@microsoft.com

ABSTRACT

Many information retrieval tasks involve searching on behalf of others. Example scenarios include searching for a present to give a friend, trying to find "cool" clothes for a teenage child, looking for medical supplies for an elderly relative [1], or planning a group activity that many friends will enjoy. In this paper, we use demographically annotated web search logs to present a large-scale study of such "on behalf of" searches. We develop an exploratory technique for recognizing such searches, and present information to describe and understand the phenomenon, including the demographics of who is searching, who they are searching for and on what topics.

1. INTRODUCTION

Many information retrieval tasks involve searching on behalf of others. Common scenarios include searching for a perfect gift, trying to find "cool" clothes for a teenage child, looking for medical supplies for an elderly relative [1] or planning a group activity that many friends will enjoy. Unlike tasks performed on behalf of oneself, information retrieval tasks performed on behalf of others are prone to two key challenges: First, the searcher is often unfamiliar with the search domain; and secondly, the searcher is often unfamiliar with the mindset of the search *beneficiary*. For example, when searching for a perfect gift for a friend, the searcher might know that the friend loves to cook, but be unfamiliar with the vocabulary used to refer to kitchen gadgets, cookbooks, and ingredients, and moreover, the searcher might not know what kinds of accoutrements the friend might appreciate. The result is an extended exploratory phase of search, where the searcher attempts to better understand both the kinds of results that can be found, as well as the likely mindset of the beneficiary.

In this paper, we present our work on quantitatively characterizing the phenomena of "on-behalf-of" searches. Using demographically identified search query data from comScore, we identify search queries where a searcher of one demographic group is performing a task on behalf of a different demographic group. We do so by focusing our task on queries that include a reference to a demographic group (e.g., [`gifts for men over 50yrs`]) and identifying instances when the referenced group differs from the searcher's own demographic group, as identified in the comScore metadata. Our primary goals in this analysis are to characterize the types of relationships between the searcher and beneficiary and identify the most frequent topics where on-behalf-of searches are observed in the logs. A secondary goal is to use these log-driven insights to reflect on how search could be altered to better support on-behalf-of searches.

As for the first question of who, demographically speaking, is searching for whom, we use the search logs to find a reflection of intuitive social patterns in our data. These patterns include patterns consistent with significant-other/marriage relationships (men perform on-behalf-of searches for women of their same age or slightly younger). We find that while some of these relationships are gendered (young men are more likely to search on behalf of men their grandfather's age; while it is young women who search on behalf of women their grandmother's age); other relationship are not (e.g. across all age ranges each gender shows a propensity to search on-behalf-of the opposite gender in the same age range as the searcher – quite possibly a spouse or significant other). Moreover, some relationships are symmetric (e.g. mothers are likely to search on-behalf-of their children; and the children are as likely to search on-behalf of their mothers).

Secondly, we characterize the topics and sub-topics for which people search "on-behalf" of others. We find that the most common topics are broadly games, clothing, and, broadly speaking, fashion and taste related. A key commonality of these topics is that individual and demographic preferences vary greatly---e.g., some brands are very popular among young adults, while others are popular among older demographics.

While this analysis method faces challenges, including ambiguities inherent in short, noisy query text, as well as potential systematic biases when searchers do not indicate a beneficiary's demographics, we believe that the result is still useful as a first large-scale identification and characterization of the phenomenon of on-behalf-of tasks.

Our findings highlight important challenges in supporting on-behalf-of searches. In some sense, it is an extreme form of exploratory search, where a searcher is attempting to understand not only the domain, resolve vocabulary mismatches, etc., but also attempting to learn the beneficiary's likely preferences for what a "good result" might be.

These observations suggest an opportunity to provide special support for on-behalf-of searches based on the aggregate search behaviors of others in the beneficiary's demographic groups. We refer to tailoring results to a search beneficiary's demographic as ***persona***-ization – essentially because the searcher can specify a persona which describes the beneficiary who may be a different person than the user but more generally could be used to describe different aspects of the user (e.g. "my business identity" vs. "my social side"). We present a brief discussion of how the aggregate behaviors of searchers within a beneficiary's demographics might be used to persona-ize the experience for on-behalf-of searchers.

Presented at SIGIR SPS 2015 Workshop, 2015.

[1] Work performed at Microsoft Research.

2. RELATED WORK

Previously, Becker *et al*. [1] studied and reported on the frequency and type of "on-behalf-of" searches in a public library setting in the United States. In particular they report that 63% of users (48.6 million people) use computers in public libraries to seek information or perform a task on behalf of others. Surveys and interviews revealed that this covers a broad range of tasks: from finding and buying car parts for friends to ordering medical supplies for elderly relatives. In the context of public libraries, the searchers tend to be from economically disadvantaged groups that may have limited alternative access points and can search on behalf of others in their community. Furthermore, there was a prevalence for searchers to be younger than the beneficiaries of the information/tasks – indicating that familiarity with technology may play a role. While search within a public library setting on behalf of others is important, we wish to determine what evidence exists to support that users search on behalf of others more generally. Research in the area of collaborative information seeking highlights challenges of explicit or implicit collaboration among two or more people, including through collaborative filtering and recommender systems [8]. While sharing many challenges, including mediation and mitigation of variances in intents and expertise, on-behalf-of search tasks differ in that the beneficiary of a search may not be a party to the search task.

A variety of work has studied both extracting profiles for personalization as well as how to tailor search results given such profiles. A sampling of such work includes extracting short- and long-term topical profiles [7][5][12][10][14], using weighted term-vectors based on long-term desktop activities [13], leveraging a user's location [2], and using the user's previous queries and clicks [4][3][11]. However, personalization makes the assumption that the user is searching for themselves with their own preferences in mind. In the setting of persona-ization, this fundamental assumption is challenged, and the need is highlighted for methods that either infer the persona (e.g. from the query) or perform a tailored search where the user is given the option of specifying the persona (e.g. query = [`video games`] where persona = "8 year old boy who loves Minecraft and Peggle").

A distinct but related area of research is separating a log from a shared device into streams associated with the distinct users in a house. For example White *et al*. [15] attempt to separate search activity on shared devices and Luo *et al*. [9] seek to differentiate user-viewing of television programs in a household. These studies may be useful in identifying the correct persona to use for tailoring search and recommendation, but in those studies, the user actively using the device is still consuming for themselves. In our setting, we target cases where the user is actively searching on behalf of someone else.

In particular, in contrast to previous work, our focus is on using large-scale search logs to analyze the type and frequency of on-behalf-of search. Furthermore, we provide motivational evidence and a description of one approach to persona-ization that leverages an interface where a user can explicitly specify a persona to conduct a search based on a different user's demographics.

3. SEARCH ON BEHALF OF OTHERS

Informally, a *search on behalf of others* is when a person is attempting to gather information to help satisfy a need or perceived need of someone other than themselves. For clarity, let us define the *searcher* as the person issuing the queries to the search engine, and the *beneficiary* as the person on whose behalf the searcher is searching.

Gift giving, purchasing or planning on behalf of another (including personal assistants), seeking medical information for another, and planning group activities are all example scenarios where an individual must search for information that will ultimately be used to satisfy someone else's needs or desires. For the searcher, there are many potential challenges when searching for another. For example, the searcher might not have full information about their target person's needs or desires. The searcher might not fully understand the given topic or domain, and may not even know the necessary vocabulary.

Our characterization of on-behalf-of demographics relies on identifying queries that specify a persona, extracting the topic of the task, identifying the demographics of the searchers from user information sources and extracting the demographics of the beneficiaries from the queries.

3.1 Identifying Queries Specifying a Persona

In this section, we describe a simple way to identify queries that explicitly specify a persona. We use a sample of one month of Bing query logs from May 2013. To reduce linguistic variation due to language and geography, we use only queries from the English US market. From these logs, we extract queries that contain one of the phrases "for women", "for men", "for boys", or "for girls" and then remove any query whose primary topic is identified as adult, pornography, or romance related. We consider the remainder to be queries that specify a persona. These contain such queries as `[birthday party ideas for boys in their teens]` and `[math games for girls]`. Overall, we find the frequency of such explicitly-specified queries to be 0.10% of all queries. Note that this is likely an underestimate of the number of both searches that have a persona in mind as well as searches on behalf since it requires matches to a fairly strict phrasing. Search interfaces that enable directly entering and leveraging a user-specified persona, likely would see a greater frequency of use. Nonetheless, given the large number of matches to even this strict phrasing, we focus on analyzing this subset as an initial foray into this domain.

Table 1: Top 20 queries that frequently contain gender/age targets

Topics	Relative % of Persona Queries
games	22.00%
shoes	10.07%
hairstyles	6.42%
tattoos	4.48%
dresses	3.74%
clothing	2.60%
shirts	2.41%
names	2.14%
sandals	2.00%
watches	1.96%
boots	1.95%
gifts	1.83%
haircuts	1.66%
clothes	1.53%
ideas	1.50%
pants	1.44%
swimsuits	1.44%
hats	1.25%
shorts	1.20%
suits	1.18%

In later sections, we will use demographics data to identify whether the persona-specified does or does not match the searcher. However, here we seek to answer the question: what topics searchers are looking for when they explicitly specify a persona?

Table 1 shows the top topics that frequently contain a persona specification. We identify topic by simply extracting the word that precedes the persona specifier (the phrases used above like "for men"). A more general future direction is not only more robustly identifying persona specifications but identifying and normalizing for topic in a more general fashion. Using this simple identification of topic, we find that the most common topics are broadly games, clothing, and, broadly speaking, fashion and taste related. A key commonality of such topics is that individual and demographic preferences vary greatly within these domains---e.g., some brands are very popular among young adults, while others are popular among older demographics. Presumably searchers believe that by explicitly specifying the demographics of the beneficiary, they are more likely to receive appropriately tailored search results.

3.2 Data and Searcher Demographics

To separate searches where the persona refers to the searcher versus a beneficiary other than the searcher, we require both the searcher's demographics and the beneficiary's demographics. To this end, we use search logs available from the internet analytics company comScore (comScore.com) with a paid subscription. We use comScore data gathered during January to February 2014 in the English-speaking segment of the US market.

Households opt-in to provide traces of their online activities, including searches at major search engines. Additionally, each person has a unique person identifier that should be used to sign-in before a user in the household uses a particular device for search. ComScore intentionally tries to construct an overall sample that is representative in the target market [6] across all major search providers. Users provide individual demographic data associated with the person identifier. This information provides us with both queries and demographic information about the searcher. Individuals are aware, of course, of comScore's data collection and are compensated in exchange.

3.3 Beneficiary Demographics

When searching for others, people often embed specific demographic attributes in their query formulations. Knowing the demographic attributes of the searcher, we can recognize when they are searching for another by looking for demographic attributes that do not match their own. For example:

- `[shoes]` vs. `[shoes for women]`
- `[workouts]` vs. `[workouts for men over 50]`

To infer demographic information about the beneficiary of a query, we extract the gender and age from query text.

In order to accomplish this, we use a small set of rules similar to those in Section 3.1. However, we broaden the set for gender to not necessarily require the leading word "for" as well as include several other gendered expressions, e.g. ("dad", "mother"). For age, we similarly recognize a small number of patterns denoting age (e.g. "NUMBER yo", "over NUMBER", etc.). Challenges are primarily due to recognizing and filtering ambiguous uses of numbers when they are not age-related, ("shoes for women under 20" often refers to the price of the shoe, not the age; "dresses for women under 100" refers to weight or price). Although our patterns recognize cases where a trailing word disambiguates (e.g. we recognized that "shoes for women over 20 dollars" is a price). Again more general query parsing to extract age and gender is an interesting area for future study. We focus on a narrow set that can be extracted relatively confidently.

If both the age and gender of the beneficiary can be extracted from the query, we assume the beneficiary is the same as the searcher if they have a non-empty intersection (e.g. a 54 year-old man searching for `[workouts for men over 50]` is assumed to be searching for himself). If we cannot accurately extract both age and gender from the query, then we assume the searcher is searching for themselves to take a conservative analysis point with respect to the phenomenon of on-behalf-of searches. Extending the dataset under consideration (e.g., through session-level or task-level identification of beneficiary demographics; or through other inference methods) remains future work.

4. CHARACTERIZING ON BEHALF OF SEARCHES

To better understand on-behalf-of searches, we wish to characterize the relationship between searchers and beneficiaries as well as understand case studies of breakdowns on some of the most common topics of on-behalf-of searches that were discussed in Section 3.1.

4.1 Demographics of "on behalf of searches"

Table 2 and Table 3 show who is searching for whom. The gender and age of the searcher are shown on the x-axis, and for Table 3, the gender and age of the beneficiary, extracted from the query, are shown on the y-axis. For Table 2, the first column is the percentage relative to any persona-specified query that is issued by that demographic (i.e. the column sums to 100%). The remaining cells are the breakdown within the searcher demographic (i.e. the second and third percentage in a row sums to 100%).

We see that when users specify a persona, they most commonly specify a persona that does not match their own demographics – with nearly all searcher demographics searching on-behalf-of others more than half of the time, with males searching for others more frequently across comparable ages, as compared to females.

Searcher		% Persona-Specified	Self	Others
Female	**18-20**	8.27%	36%	64%
	21-24	9.34%	6%	94%
	25-34	10.07%	31%	69%
	35-44	8.08%	30%	70%
	45-49	3.40%	39%	61%
	50-54	2.94%	20%	80%
	55-64	3.44%	53%	47%
	65+	1.54%	24%	76%
Male	**18-20**	8.54%	11%	89%
	21-24	12.76%	9%	91%
	25-34	13.33%	7%	93%
	35-44	6.63%	22%	78%
	45-49	3.27%	8%	92%
	50-54	2.81%	15%	85%
	55-64	3.42%	10%	90%
	65+	2.17%	2%	98%

Table 2: Percentage of all queries specifying a persona by searcher demographic and breakdown into self-search and on-behalf search.

Searcher		Beneficiary: Female 18-20	21-24	25-34	35-44	45-49	50-54	55-64	65+	Male 18-20	21-24	25-34	35-44	45-49	50-54	55-64	65+	
Female	18-20	36%	2%	9%	5%	7%	6%	5%	0%	10%	5%	4%	2%	1%	4%	3%	1%	18-20
	21-24	15%	6%	10%	8%	22%	1%	14%	5%	6%	3%	1%	5%	0%	0%	3%	1%	21-24
	25-34	5%	4%	31%	11%	4%	12%	9%	4%	2%	1%	4%	1%	2%	2%	4%	4%	25-34
	35-44	10%	3%	3%	30%	13%	0%	5%	7%	4%	0%	1%	9%	6%	3%	4%	1%	35-44
	45-49	3%	0%	1%	13%	39%	9%	9%	0%	12%	0%	0%	4%	4%	2%	3%	2%	45-49
	50-54	2%	0%	27%	3%	4%	20%	28%	3%	2%	0%	0%	0%	4%	1%	2%	3%	50-54
	55-64	1%	0%	3%	3%	2%	10%	53%	17%	0%	0%	1%	1%	0%	3%	6%	1%	55-64
	65+	0%	0%	0%	2%	8%	24%	35%	24%	0%	0%	0%	1%	0%	0%	2%	2%	65+
Male	18-20	34%	4%	4%	9%	7%	8%	3%	1%	11%	0%	4%	5%	6%	0%	2%	1%	18-20
	21-24	26%	4%	4%	7%	3%	6%	5%	2%	9%	9%	4%	4%	2%	0%	15%	0%	21-24
	25-34	19%	5%	9%	9%	9%	5%	11%	5%	8%	1%	7%	4%	1%	3%	4%	1%	25-34
	35-44	17%	1%	0%	14%	8%	2%	10%	8%	2%	0%	7%	22%	8%	0%	1%	1%	35-44
	45-49	16%	1%	6%	16%	20%	1%	4%	12%	0%	0%	0%	8%	8%	5%	0%	5%	45-49
	50-54	16%	1%	12%	7%	22%	7%	11%	0%	2%	1%	0%	0%	2%	15%	3%	1%	50-54
	55-64	15%	0%	6%	14%	13%	8%	21%	6%	1%	2%	0%	0%	0%	3%	10%	1%	55-64
	65+	4%	0%	2%	10%	14%	4%	14%	32%	3%	0%	1%	1%	3%	10%	1%	2%	65+

Table 3: Demographics of beneficiary for specified persona in query relative to searcher's demographics

Searcher		Beneficiary: Female 18-20	21-24	25-34	35-44	45-49	50-54	55-64	65+	Male 18-20	21-24	25-34	35-44	45-49	50-54	55-64	65+	
Female	18-20	73%	4%	0%	0%	0%	0%	5%	0%	18%	0%	0%	0%	0%	0%	0%	0%	18-20
	21-24	52%	0%	0%	0%	16%	0%	30%	0%	2%	0%	0%	0%	0%	0%	0%	0%	21-24
	25-34	13%	0%	52%	0%	0%	9%	0%	17%	0%	9%	0%	0%	0%	0%	0%	0%	25-34
	35-44	0%	0%	0%	0%	20%	0%	12%	0%	68%	0%	0%	0%	0%	0%	0%	0%	35-44
	45-49	0%	0%	0%	22%	42%	0%	32%	0%	4%	0%	0%	0%	0%	0%	0%	0%	45-49
	50-54	2%	0%	79%	1%	5%	5%	8%	0%	0%	0%	0%	0%	0%	0%	0%	0%	50-54
	55-64	0%	0%	0%	12%	0%	0%	62%	26%	0%	0%	0%	0%	0%	0%	0%	0%	55-64
	65+	0%	0%	0%	0%	0%	0%	18%	82%	0%	0%	0%	0%	0%	0%	0%	0%	65+
Male	18-20	24%	0%	20%	0%	0%	0%	0%	0%	13%	0%	0%	2%	17%	0%	24%	0%	18-20
	21-24	8%	0%	20%	0%	13%	0%	0%	0%	9%	0%	0%	5%	42%	0%	2%	0%	21-24
	25-34	4%	0%	0%	20%	52%	0%	0%	0%	18%	0%	0%	0%	6%	0%	0%	0%	25-34
	35-44	0%	0%	0%	42%	0%	0%	15%	0%	0%	0%	0%	8%	0%	0%	34%	0%	35-44
	45-49	0%	0%	10%	0%	46%	0%	0%	44%	0%	0%	0%	0%	0%	0%	0%	0%	45-49
	50-54																	50-54
	55-64	0%	0%	0%	0%	0%	3%	0%	0%	0%	0%	0%	0%	0%	0%	97%	0%	55-64
	65+	0%	0%	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%	0%	0%	0%	65+

Table 4: In "clothes" category, demographics of beneficiary relative to searcher's demographics

From the detailed breakdown in Table 3, we see several patterns of searches reflecting intuitive relationships among demographic groups:

- People likely searching on-behalf-of their children (e.g. 27% of persona-specified queries for women ages 50-54 have a beneficiary of women age 25-34; 12% of persona-specified queries for men ages 50-54 have a beneficiary of men age 25-34). E.g., [`christmas gifts for a 25 year old child`]
- People likely searching on-behalf-of their parents and grandparents (e.g. 15% of persona-specified queries for men ages 21-24 have a beneficiary of women age 55-64; 5% of persona-specified queries for women ages 25-34 have a beneficiary of women age 65+). E.g., [`present for grandma 60 years`]
- People searching on-behalf-of their spouses/significant others: note the significant diagonal in the *opposite* gender from the searcher in the same age category. E.g., [`birthday gifts 30 year old female`]

4.2 Demographics and Topics of "on behalf of searches"

We can also consider certain topical slices. Table 4 presents a similar table to Table 3, but restricted to queries that mention clothes. Note that some cells are missing here due to sparsity given the sample size.

Looking at the relationship between demographics and topics, we see that many of the symmetric relationships we identified in Section 4.1 seem to lose their symmetry. For example, while mothers look for clothes on behalf of their children, their children do not necessarily search for clothes on behalf of their parents.

5. DISCUSSION

5.1 Using Self-Searches to Aid On-Behalf-Of Searches

In Table 5, we can see how style- and brand-popularity varies significantly across gender and age groups. Only a small number of keywords (`Jordan` and `Nike`) are popular across most age groups and both genders. Some (for example, `Vans`, `Converse`) are popular only among a single demographic. Other observations include that at all comparable ages, a larger proportion of males'

	Searcher: Female								Searcher: Male							
	18-20	21-24	25-34	35-44	45-49	50-54	55-64	65+	18-20	21-24	25-34	35-44	45-49	50-54	55-64	65+
adidas	4%	4%	3%	3%	2%	3%	2%	4%	6%	4%	4%	5%	3%	4%	2%	2%
bakers	2%	3%	3%	2%	2%	2%	2%	1%	1%	1%	1%	1%	1%	1%	1%	1%
basketball	5%	5%	4%	7%	8%	4%	5%	3%	9%	11%	8%	10%	10%	7%	5%	8%
boat	2%	2%	2%	2%	1%	2%	1%	2%	3%	3%	4%	1%	2%	3%	3%	1%
cheap	3%	4%	5%	2%	3%	1%	1%	0%	2%	2%	3%	2%	1%	2%	1%	1%
converse	3%	3%	3%	3%	2%	2%	1%	2%	2%	2%	2%	2%	3%	2%	1%	1%
dance	2%	2%	2%	2%	2%	3%	4%	5%	1%	0%	1%	1%	1%	1%	1%	5%
dc	2%	2%	2%	2%	2%	3%	1%	1%	2%	3%	2%	2%	1%	1%	1%	0%
dress	3%	4%	3%	4%	4%	6%	8%	10%	3%	3%	4%	5%	6%	5%	5%	7%
golf	0%	0%	1%	1%	1%	1%	2%	6%	1%	1%	3%	3%	3%	6%	8%	13%
gucci	1%	1%	1%	1%	1%	1%	1%	1%	2%	2%	2%	2%	1%	1%	0%	0%
heel	3%	3%	5%	2%	2%	1%	2%	3%	0%	1%	1%	1%	1%	1%	1%	1%
james	2%	2%	2%	2%	2%	1%	1%	0%	3%	3%	2%	3%	3%	1%	1%	1%
jordan	14%	12%	10%	10%	9%	8%	8%	4%	18%	15%	14%	11%	7%	8%	7%	7%
jordans	3%	3%	2%	2%	2%	1%	1%	1%	3%	3%	2%	2%	1%	1%	1%	1%
mens	1%	1%	2%	3%	2%	2%	2%	4%	2%	2%	3%	2%	3%	4%	6%	5%
nike	9%	8%	7%	7%	6%	6%	4%	2%	10%	8%	9%	9%	10%	5%	4%	5%
payless	6%	7%	10%	10%	14%	19%	17%	15%	2%	2%	3%	6%	11%	10%	9%	8%
room	3%	3%	3%	4%	3%	3%	4%	4%	1%	1%	1%	2%	2%	4%	3%	3%
running	6%	6%	6%	7%	9%	7%	5%	5%	7%	8%	11%	9%	10%	13%	17%	11%
skate	2%	3%	1%	2%	2%	2%	1%	1%	3%	4%	3%	2%	2%	2%	1%	1%
soccer	3%	2%	3%	2%	1%	1%	1%	1%	5%	5%	4%	4%	5%	2%	1%	1%
supra	2%	1%	1%	2%	2%	1%	0%	0%	4%	4%	2%	3%	1%	2%	2%	3%
tennis	4%	4%	4%	5%	6%	6%	7%	9%	2%	2%	4%	5%	4%	5%	10%	3%
vans	3%	3%	1%	2%	1%	2%	2%	1%	3%	3%	2%	1%	2%	1%	1%	2%
wedding	4%	4%	6%	2%	3%	2%	3%	2%	1%	1%	2%	1%	1%	2%	1%	1%
women	2%	3%	4%	2%	2%	3%	4%	5%	0%	0%	0%	0%	0%	0%	0%	0%
womens	2%	2%	3%	2%	3%	4%	6%	7%	1%	1%	1%	1%	2%	2%	3%	5%
wrestling	1%	0%	1%	2%	2%	2%	0%	1%	2%	2%	2%	3%	2%	2%	1%	1%
your	2%	1%	1%	2%	1%	2%	1%	1%	2%	3%	2%	1%	2%	2%	2%	1%

Table 5: Relative popularity of keywords used, $P(keyword \mid age, gender)$, when searching for "shoes", by demographic.

shoe queries are related to golf and running shoes as compared to females. Likewise, among males, the proportion of shoe queries related to basketball, Jordan, Nike and Adidas shoes decreases with increasing age, while queries related to dress, golf, and running shoes increase with age. These types of trends are not obvious to any searcher who wishes to search on behalf-of-another demographic. By identifying the prevalent particular items within a targeted category, search data can be used to improve a searcher's exploration by exposing to them – either indirectly through re-ranked results or directly through suggested queries/items – the behavior of the demographic they wish to search on-behalf-of.

5.2 Beyond Demographics

When people search on-behalf-of another, they are not limited to characterizing a beneficiary based on the demographic details. Searchers often know other key information, such as their likes and preferences and embed this information into queries (`[games for people who like star wars]`). In the same way that we used age and gender information embedded in a query to identify beneficiary demographics, we can use these hints to identify a cohort similar to a given beneficiary. That is identify searchers from the targeted demographic whose search history indicates a preference from star wars. That identified cohort can then be restricted to game-related queries and the resulting queries can be used to expose to the searcher games that would likely match the interest and demographics of the beneficiary.

6. SUMMARY

In this short paper, we provide the first large-scale characterization of the phenomena of on-behalf-of searches. We characterized the demographic relationships embedded within such searches, as well as the primary topics that are the subject of such searches, highlighting key challenges faced by on-behalf-of searchers when attempting to complete their tasks.

In future work, we would like to explore on-behalf-of searches at a session level, looking at metrics of satisfaction and task completion, and further exploring how the aggregate behaviors of beneficiary cohorts can be used to improve the on-behalf-of search experience.

7. REFERENCES

[1] S., Becker, M.D. Crandall, K.E. Fisher, B. Kinney, C. Landry, and A. Rocha. Opportunity for All: How the American Public Benefits from Internet Access at U.S. Libraries. (IMLS-2010-RES-01). Institute of Museum and Library Services. Washington, D.C. 2010. http://impact.ischool.washington.edu/documents/OPP4ALL_FinalReport.pdf

[2] P.N. Bennett, F. Radlinski, R.W. White, and E. Yilmaz. Inferring and Using Location Metadata to Personalize Web Search. In SIGIR '11, pp. 135–144, 2011.

[3] H. Cao, D.H. Hu, D. Shen, D. Jiang, J. Sun, E. Chen, and Q. Yang. Context-aware Query Classification. In SIGIR '09, pp. 3–10, 2009.

[4] H. Cao, D. Jiang, J. Pei, Q. He, Z. Liao, E. Chen, and H. Li. Context-aware query suggestion by mining click-through and session data. In KDD '08, pp. 875–883, 2008.

[5] P.A. Chirita, W. Nejdl, R. Paiu, and C. Kohlschütter. Using ODP Metadata to Personalize Search. In SIGIR '05, pp. 178–185, 2005.

[6] Fulgoni, G.M. The "Professional Respondent" Problem in Online Survey Panels Today. Slides online at: http://www.sigmavalidation.com/tips/05_06_02_Online_Survey_Panels.ppt (Downloaded on June 1, 2015). 2005.

[7] S. Gauch, J. Chaffee, and A. Pretschner. Ontology-Based User Profiles for Search and Browsing. Web Intelligence and Agent Systems 1, 3-4, pp. 219–234, 2004.

[8] G. Golovchinsky, P. Qvarfordt, J. Pickens. Collaborative Information Seeking. IEEE Computer, 42(3): 47-51, 2009.

[9] D. Luo, H. Xu, Hongyuan Zha, J. Du, R. Xie, X. Yang, and W. Zhang. You Are What You Watch and When You Watch: Inferring Household Structures from IPTV Viewing Data. IEEE Transactions on Broadcasting Technology, 60(1): 61-72, 2014.

[10] Z. Ma, G. Pant, and O.R.L. Sheng. Interest-based Personalized Search. TOIS (25:1), Article 5, 2007.

[11] L. Mihalkova and R. Mooney. Learning to Disambiguate Search Queries from Short Sessions. In ECML PKDD '09, pp. 111–127, 2009.

[12] M. Speretta and S. Gauch. Personalized search based on user search histories. In Web Intelligence '05, pp. 622–628, 2005.

[13] J. Teevan, S.T. Dumais, and E. Horvitz. Personalizing Search via Automated Analysis of Interests and Activities. In SIGIR '05, pp. 449–456, 2005.

[14] R.W. White, P.N. Bennett, and S.T. Dumais. Predicting short-term interests using activity-based search context. In CIKM '10, pp. 1009–1018, 2010.

[15] R.W. White, A. Hassan, A. Singla, and E. Horvitz. From Devices to People: Attribution of Search Activity in Multi-User Settings. In WWW '14, pp. 431-442, 2014.