DOI: http://dx.doi.org/10.14236/ewic/FDIA2015.17

Query Recommendation as Query Generation

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

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

Keywords: Query recommendation, Information retrieval, IR, Exploratory search, Recommender systems, Web crawl

1. INTRODUCTION

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

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

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

2. BACKGROUND

2.1. Query Recommendation

Query recommendation has a long history dating to at least the early 2000s, yielding various techniques with differing applications. Most approaches harness some aspect of search logs to make personalised recommendations to users. These techniques can utilise sequences of queries in a log, click-through information on SERP URLs, the snippets in SERP results and the full text of linked pages. Baeza-Yates et al. (2004), for instance, use click-through information to cluster queries by similarity, offering similar substitutes when a current query fails to yield good results. An opposing application from Vahabi et al. (2013) similarly uses query distance to recommend orthogonal queries - dissimilar queries that are similar enough to be topically relevant. He et al. (2009) and Boldi et al. (2008) model n-grams and even general query graphs respectively, lending themselves to the applications in the Introduction.

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

2.2. Modeling the Web

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

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

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

2.3. Exploratory Search

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

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

3. METHODOLOGY

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

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

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

4. EVALUATION

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

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

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

5. ACKNOWLEDGMENTS

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

REFERENCES

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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