Query Recommendation for Children

Sergio Duarte Torres and Djoerd Hiemstra
University of Twente
The Netherlands
duartes,hiemstra@cs.utwente.nl

Ingmar Weber
Yahoo! Research
Spain
ingmar@yahoo-inc.com

Pavel Serdyukov
Yandex
Russia
pavser@yandex-team.ru

ABSTRACT

One of the biggest problems that children experience while searching the web occurs during the query formulation process. Children have been found to struggle formulating queries based on keywords given their limited vocabulary and their difficulty to choose the right keywords.

In this work we propose a method that utilizes tags from social media to suggest queries related to children topics. Concretely we propose a simple yet effective approach to bias a random walk defined on a bipartite graph of web resources and tags through keywords that are more commonly used to describe resources for children.

We evaluate our method using a large query log sample of queries aimed at retrieving information for children. We show that our method outperforms query suggestions of state-of-the-art search engines and state-of-the art query suggestions based on random walks.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Query formulation

General Terms

Experimentation, Measurement

Keywords

query formulation, children, social media

1. INTRODUCTION

Children experience several difficulties searching the web using state-of-the-art search engines. In particular, children have been found to struggle formulating queries with keywords [2, 6].

In this work we propose a query suggestion method to help children find keywords that are more likely to be relevant for them. Although several research has been carried out on query suggestion[15, 17], our work deviates from previous studies in that *(1)* the suggestions are aimed at potential children search intents; *(2)* the suggestions are constructed in the absence of query logs and *(3)* the suggestions are ranked based on a novel biased random walk to promote suggestions aimed at children topics.

The query suggestions of our method are based on the tags from the bookmarking system *Delicious* that are associated to the query web results and to previously seen web resources intended for children. These type of tags are a valuable resource in the domain of IR for children since we can exploit the collaborative information provided by users sharing web resources for children.

Concretely, we propose a novel way to boost tags in a random walk that are more frequently used to describe resources for children and that are more prominent with respect to a background model of web resources aimed at the general public. The model is built using a subset of the large *delicious* crawl described in [16]. The subset consists of bookmarks of high quality of web resources focused on content for children. The assumption of our method is that tags more frequently associated to urls focused on children topics are better candidates to construct query suggestions for children. For instance consider the query *cars*. According to Google's query suggestions common aspects associated to this query are *car rentals, cars for sale, used cars, new cars, disney cars* and *car pictures*. On the other hand, aspects oriented to satisfy children information needs should rather include aspects as *car games, car toys, car movies, car images, car colouring pages*. Our system ranks higher the latter tags providing suggestions more focused on content for children.

The quality of the results is evaluated using a large log sample of queries and query reformulations that are aimed at retrieving information for children. The data set was extracted from the AOL query log and we utilize queries landing on domains listed in the Kids and Teens directory of Dmoz.

The organization of this paper is as follow: Section 2 describes the most relevant related work to this paper, where we emphasize query expansion methods and tag ranking. Section 3 describes our method employed. Section 4 describes the data acquisition process. Section 5 presents the results obtained by our random walk method. In the last section conclusions of this work and some directions for future work are discussed.

CIKM'12, October 29–November 2, 2012, Maui, HI, USA.

2. RELATED WORK

This work is related to the areas of query suggestion, query expansion and tag ranking. In the following paragraphs we review the most relevant research related to our work.

2.1 Query Recommendation

Extensive research has been carried out on query recommendation based on query click-through data from query logs [8, 1]. In these methods the association between query and documents in the search graph is mined to infer related queries. More recently, random walk frameworks have been proposed to rank documents and queries using hitting time [12] and based on the query - document frequency in the graph[4, 5].

The method we proposed utilized Craswell et al.'s [5] framework. However, our work deviates from theirs in the definition of the transition probabilities and the normalization of these probabilities. Our motivation is to bias the walk towards suggestions more appropriate for a specific niche of users (i.e., children) which is not addressed by their work.

Recently, two random walk frameworks have been proposed to leverage query logs and social media annotation within the same graph. The first exploits the latent topic space of the graph [3] and the second utilizes the graph hitting time [12]. Their focus is to refine queries by exploiting the tag vocabulary of the social media and to provide exploratory and search query suggestions within the same framework. Our work also exploits the annotations of social media to the generation of query suggestions however our work differs from theirs in that our method provides query suggestions in the absence of query logs, that is solely based on social media.

2.2 IR for Children

Gyllstrom et al. [9] presented a variation of Page Rank to rank web pages that are more suitable for children. They utilized label propagation to score the documents. Our work deviates from theirs in the characteristics of the graph utilized (we employ social media while they employ web documents) and in that we boost query suggestions associated with content for children through information metrics between a foreground model of tags used to describe content for children and a background model. Moreover we improve over the results of our random walk using a learning to rank framework by introducing features that are not trivial to add in a graph based model.

2.3 Tag Ranking

Tag ranking has recently received attention given the proliferation of social media sharing sites. Liu et al. [10] proposed a method to estimate the relevance score of a tag to an image based on probability density estimation. The estimation is further refined using a random walk over a tag similarity graph. Our work deviates from theirs in the structure of the graph and the bias introduced into the random walk. In [10] the graph consists only of tags and in our problem the graph consists of tags and web resources. This graph structure is important in our problem since we exploit the characteristics of web resources aimed at children to bias the random walk.

3. METHOD

In this section we describe the scenario in which query expansion is studied and the method proposed to mine query suggestions using tags from social media.

3.1 Problem Scenario

We envisage a search service for children which reuses state-of-the art search engines to deliver content aimed at children. In this system, the query submitted by the user is sent to several search engines to retrieve keywords from the snippets and titles of the web results. These keywords represent the possible topics associated to the user's query. Our task is to generate these keywords and rank them to construct query suggestions. Note that in this scenario we do not have access to search engine query logs which are widely used for query recommendation [4, 11]. Moreover given the increasing concern of users for their privacy and the characteristics of the audience targeted by our system (i.e. children), it is desirable to avoid the tracking of user information.

3.2 Random Walk Towards Content for Children

Our random walk model uses a bipartite graph of web resources (i.e., *urls*) and tag nodes. Previous research on tag ranking [10] employed random walks methods for tag recommendation systems using a graph composed solely of tags. In the setting of our problem we found it useful to treat urls as nodes as well since our methods rely on a trusted set of web resources, which are used as seeds to bias the random walk towards more relevant tags for the targeted audience. That is, tags more frequently associated to urls that are known to be targeted at a certain niche of users (i.e. children) will be promoted over tags employed more frequently to described urls for a different niche of users (i.e. adults). Note that is not straight forward to represent this information in the case where the graph is only composed of tag nodes, moreover this graph representation allows to add a measure of how reliable or trustful a seed url is (e.g., based on its source or popularity).

In this work the graph was created using a set of the *Del.icio.us* bookmarks from the Wetzker et al. [16] collection. Concretely, bookmarks of urls known to be adequate for children were extracted to create the set of urls and tags. Details about the characteristics of the dataset are provided in Section 4.1. Our random walk method is based on the framework proposed by Craswell and Szummer [5]. Formally the graph is defined as:

Definition 1. (Bipartite graph) *a bipartite graph of urls and tags :*
$G = (U, T, E = \{(u,t)|(u,t) \in U \times T\})$
where $U = \{u_1, u_2, .., u_n\}$ *is the set of urls described by tags* $T = \{t_1, t_2, .., t_m\}$ *and* E *is the set of edges in the graph.*

In [5] the transition probabilities are defined as:

$$p_{fw}(i|j) = \begin{Bmatrix} (1-\alpha)\frac{c(i,j)}{\sum_{k:(j,k)\in E} c(j,k)} & \textit{if } i \neq j \\ \alpha & \textit{for } i = j \end{Bmatrix} \quad (1)$$

The term $c(i,j)$ represents the number of times a tag i was used to describe a web resource j and the term α is the self transition probability which is used to slow the diffusion of the scores. We employed this weighting scheme as baseline for our method.

We propose to bias the random walk by introducing a weight based on the point-wise Kullback-Leibler (KL) divergence metric. Intuitively, this metric allows in a straight forward manner to promote those tags that have a greater expectation to appear in a collection of content for children (our foreground model) than in a corpus of content for grown ups (background model). Equation 2 and 3 reflect the new transition functions.

$$p_{fwKL}(i|j) = p(i) log \frac{p(j)}{g(j)} p_{fw}(i|j) \tag{2}$$

$$p_{bwKL}(i|j) = \begin{Bmatrix} (1-\alpha)\frac{p_{fwKL}(j|i)}{\sum_{k:(i,k)\in E} p_{fwKL}(k|i)} & if\ i \neq j \\ \alpha & if\ i = j \end{Bmatrix} \tag{3}$$

where $p(i)$ is the probability of a tag (or url) to appear in the collection of resources for children and $g(j)$ is the probability of i to appear in the collection of resources for the general public. We normalize the point-wise Kullback-Leibler (KL) distances to lie between 0 and 1 in order to introduce them into the random walk framework. The normalization was carried out using the maximum and minimum KL point-wise distance in the collection in the following manner: $kl_n(p||q) = kl(p||q) - minKL/(maxKL - minKL)$.

We also found that using a uniform normalization for the transition of *urls* to *tags* improves the performance of the random walk. Intuitively, this occurs because the standard transitions of urls to tags tend to promote the most popular tags, however our focus is to promote those tags that are more children oriented, which are not necessarily the most popular for a given url. Thus, a uniform normalization emphasizes the effect of the KL weight introduced in Equation 2 and 3. Using this observation we renormalized the forward probability as follows:

$$p_{fwN}(i|j) = \begin{Bmatrix} (1-\alpha)\frac{c(i,j)}{\sum_{k:(j,k)\in E} c(j,k)} & if\ i \neq j,\ j \in T \\ (1-\alpha)\frac{p_{fw}(j|i)}{\sum_{i:(j,i)\in E} p_{fw}(j|i)} & if\ i \neq j,\ j \in U \\ \alpha & if\ i = j \end{Bmatrix} \tag{4}$$

From Equation 2 we need to estimate the probabilities of the tags and urls in the two corpora. These probabilities are estimated based on a set of Delicious bookmarks that represent the interests of the target group.

We define a bookmark as a tuple containing a url and a tag, which describes the url: $b = \langle u_i, t_i \rangle$ where $u_i \in U, t_i \in T$, the set of urls and tags respectively. A collection of bookmarks is defined as a bag of N bookmarks $B = \{1, b_2, .., b_N\}$.

We employ a set of bookmarks that contains trusted and oriented urls for a specific target audience (i.e. children).

Definition 2. (Bookmarks for *kids*) *The bag of bookmarks of trusted and oriented urls for a target audience is defined as:*
$B_k = \{b_1, b_2, .., b_N | proj_{url}(b_i) \in U_k\}$ *where* U_k *is the set of seeds urls.*

The estimation of the transition probabilities depicted in Equation 2 is estimated using maximum-likelihood estimation (MLE) using B_k for the foreground model and B for the background model

$$\begin{aligned} p(t) = \frac{cf_{B_k}(t)}{|T|}, p(u) = \frac{cf_{B_k}(u)}{|U|} \\ g(t) = \frac{cf_B(t)}{|T|}, g(u) = \frac{cf_B(u)}{|U|} \end{aligned} \tag{5}$$

where $|T|$ and $|U|$ is the raw size of tags and urls in the collection B_k

3.3 Query representation

The query is represented as a single node in the graph and we define a special transition probability from the query node to the tag nodes of the graph. We do not include transition probabilities from the query to url nodes because the user's query is represented as a bag of tags. The query representation is constructed from the query itself and the tags found in the titles and snippets of the top ranked web results. The query can also be seen as a document constructed with the tags found in the web results and the query. Formally we define the user's query and the tag set of a query as :

Definition 3. (Query) *A query* q *of length* l *is represented as the sequence of words* $(w_1, w_2, .., w_l)$.

Definition 4. (Tag set of a query) *The tag set of a query* q *consists of the* m *tags extracted from a social bookmarking system S, which are associated to the top web results of query* q*:* $Q = \{t_1, t_2, .., t_m\}$.

This representation is convenient because query suggestions can often be obtained directly from the keywords appearing in the snippets of the web results. Using a sample of 10K queries from the AOL query log we found that the intersection between the keywords generated from the snippet/title and the vocabulary of the query reformulations (and which are also present as tags in Delicious) was 65%.

Using this query representation we define the transition probability $p(t|Q)$ as:

$$\begin{aligned} p(t|Q) &= \frac{p(Q|t)p(t)}{p(Q)} \\ p(t|Q) &\propto p(t)p(Q|t) \\ p(t|Q) &\propto p(t)\prod_{i=1}^{|Q|} p(q_i|t) \end{aligned} \tag{6}$$

The first term on the right hand side is the likelihood of the candidate tag t in the collection and the second term describes the likelihood of t co-occurring between the tags in the query and the collection. These probabilities are estimated using MLE in a similar fashion as in 5.

$$p(q_i|t) = \frac{cf(q_i, t) + \mu\, p(q_i)}{|T| + \mu} \tag{7}$$

where $p(q_i)$ is the prior probability of q_i and μ is the Dirichtlet smoothing parameter.

4. DATA SET EXTRACTION

4.1 Training Data

As training data we created a set of *Del.icio.us* bookmarks from the Wetzker et al. [16] collection. To the best of our knowledge this is the largest collection of social tagged data available for research. The collection contains 132 million bookmarks and 420 million tag assignments, and it was retrieved between December of 2007 to April of 2008. The set was created by extracting the bookmarks of the urls listed in the *Kids and Teens* section of the Open Directory Project *ODP* (only exact matches). These urls link to "web sites that have been selected for age-appropriate content by a team of volunteer editors" [1]. These resources have also been used in other information retrieval problems for young users [9, 7] with positive results. The data set was cleaned by normalizing multi-worded tags and removing ill-defined and infrequent tags (tags submitted by less than 3 three users).

4.2 Test Data

Search sessions were created using the standard 30 minutes window in order to extract query tuples. A query tuple consists of a query and a query reformulation occurring within the same search session. A query q' is a query reformulation of q if the former is a prefix (e.g. *brit, britney spears*) or a suffix of the latter (e.g. *wars cheat codes, lego star wars cheat codes*), or the latter contains all the words of the former plus another word, independently of the order in which the words appears (e.g. *york giants, super bowl york giants xxv*) and there are no query events between them. We only consider queries that land on the domains listed in the Dmoz *kids and teens* section [14] . For the evaluation we grouped query tuples according the target audience of the urls (i.e., *kids*, *teens* and *adults*). In Dmoz, urls tagged as for *kids* and *teenagers* represent urls that are appropriate for children aged, respectively, 8 to 12 and 13 to 16 years old. The query tuples targeting content for grown ups (*adults)* were created using a sample of the queries landing on the general Dmoz directory (non*kids and teens*).

Using this methodology we were able to extracted around 480K queries and 20K sessions. From these sessions we obtained in the order of tens of thousands of query pairs for each one of three age groups.

5. RANDOM WALK EVALUATION

The purpose of the evaluation is to quantify the coherence and appropriateness of the query suggestions provided by the methods described in section 3.

Assessing the quality of query suggestions can be a very hard task given that the intent of the user is rarely clear from solely the query. However, we consider that the query suggestions that are submitted by users of a given age range represent a good approximation of good query suggestions for this particular segment of users. A similar assumption has been adopted in previous query recommendation studies [13, 3].

The performance of the query recommendation task was measured in terms of recall, NDCG. All the metrics are calculated based on the set of query tuples extracted and described in the section 4.2. We employed a subset of the AOL log to tune the parameters involved in the random walk. We set the number of iterations of the random walk to *100* and we set the parameter α to *0.1*. The smoothing parameter m was set to *1200*.

To calculate the performance scores we define the set of query pairs from the gold standard as $G = \{\langle q, q'\rangle\}$ where q' is a query reformulation of q. And the set $S_n = \{\langle q, q', r\rangle\}$ where q' is a query suggestion of q and r is the ranked position of q'. For instance, recall is calculated as $r = \frac{|S_n \cap G|}{|G|}$. The intersection between the set of query suggestions and reformulations is performed using exact matching.

[1]www.dmoz.com

AOL log queries		
monsters	bing	truck games, jobs, high, jam, energy
	rw+kl	inc,music, film, pixar, images
	gold	scary
sol practice	bing	in computer, fourth grade
quizzes	rw+kl	test, learning, history, science sites
	gold	world history, history
rashes	bing	in children, pictures, of the skin, itching
	rw+kl	skin , definition, red skin, symptoms
	gold	red skin
art	bing	van, institute of chicago, institute
	rw+kl	museum, prints, school, poetry
	gold	angel food, food

Table 1: Query examples from the two query logs utilized

Table 1 presents query examples of the query suggestions provided by our method, Bing and the gold standard defined by the queries extracted from the AOL logs.

5.1 Experimental Results

We evaluate the random walk baseline (Equation 1) and our random walk method (Equation 2). Additionally, we compare the results obtained by the methods against a state-of-the-art search engine query suggestions. The graph constructed utilizes the domains from the Dmoz directory labelled as suitable for children up to 12 years old. The graph contains 91.6K edges and 20K nodes (12.9K urls and 7.1 tags). In the tables shown in this section the baseline will be referred to as *rw* and our method as *rw-kl*.

Tables 2 shows the recall values obtained for the query pairs extracted. We found that our method outperforms the baseline and the Bing query suggestions for the *children* queries and the *teenager* queries. However this is not the case for the set of *adults* queries, which was expected given that our random walk method gives priority to tags that are more popular for children.

We observed that the maximum gain obtained was for the *children* queries when considering the top 10 results: +7.5% with respect to the baseline and +9.6% with respect to Bing. For the teenagers queries the maximum gain is obtained at top 5 results: +2.3% with respect to the baseline and +2.9% with respect to Bing. Interestingly the gain observed for teenagers is not a high as the gain obtained for the *children* dataset. This may be due to the fact that the seed urls employed to build the graph are associated to web resources for children and not teenagers. In future work we will explore the possibility of building customize graph models for other age groups (e.g. teenagers).

The performance trends observed for the recall results were also reflected in the NDCG scores, as is shown in Table 3, which shows that the quality of the ranking is also improved by a reasonable margin. It is important to mention that the low values of NDCG reported are due to the sparsity of the data. On average we collected 1.6 query suggestions per query. Nonetheless, numbers on the same order have been reported on query recommendation studies for long-tailed queries [13].

query set	bing	rw	rw-kl	gain
		Top 5		
children	1.5%	3.32%	9.17%	5.85%
teenagers	2.3%	1.93%	5.25%	3.32%
adults	5.1%	0.37%	0.31%	-0.06%
		Top 10		
children	2.15	4.15%	11.71%	7.56%
teenagers	3.2%	4.2%	6.54%	2.34%
adults	5.1%	0.71%	0.71%	0.00%
		Top 50		
children	2.1%	13.2%	15.8%	2.60%
teenagers	3.2%	6.3%	9.92%	3.62%
adults	5.1%	1.0%	1.0%	0.00%

Table 2: Recall comparison across the methods using the AOL log

query set	bing	rw	rw-kl	gain
		Top 5		
children	0.017	0.021	0.069	0.048
teenagers	0.024	0.016	0.045	0.029
		Top 10		
children	0.026	0.032	0.082	0.05
teenagers	0.031	0.029	0.042	0.013
		Top 50		
children	0.026	0.076	0.089	0.013
teenagers	0.031	0.017	0.042	0.025

Table 3: NDCG comparison across the methods using the AOL search logs

6. CONCLUSIONS

In this paper we presented how tags from social bookmarking system can be exploited to produce query suggestions for a specialized group of users using a set of seed web resources and a biased random walk based on pointwise KL divergence between a foreground model and background model. Our method can be used to improve current search assistance functionality for children since we show that our method performs the best for queries aimed at the the youngest group of users (i.e. children between 10 to 12 years old). We show that our method clearly outperforms state of the art search engine query suggestions for this type of queries. We also showed that social media is a highly valuable resource for the generation of query suggestions and its use can replace the utilization of query logs which may not be available to several search systems. For future work we are interested in applying the method proposed in different domains and on different age segments. We are also interested in enriching our query suggestion method by combining it with other topical features. This can be achieved in a LTOR approach by using our method as one of the features.

7. ACKNOWLEDGEMENTS

This research is supported by the European Community's Seventh Framework Programme FP7/2007-2013 under grant agreement no. 231507, by the Torres Quevedo Program of the Spanish Ministry of Science and Innovation, by the European Social Fund, and by the Spanish Centre for the Development of Industrial Technology under the CENIT program, project CEN-20101037.

8. REFERENCES

[1] R. Baeza-Yates and A. Tiberi. Extracting semantic relations from query logs. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '07, pages 76–85, New York, NY, USA, 2007. ACM.

[2] D. Bilal. Children's use of the yahooligans! web search engine iii. cognitive and physical behaviors on fully self-generated search tasks. *J. Am. Soc. Inf. Sci. Technol.*, 53(13):1170–1183, 2002.

[3] L. Bing, W. Lam, and T.-L. Wong. Using query log and social tagging to refine queries based on latent topics. In *CIKM*, pages 583–592, 2011.

[4] P. Boldi, F. Bonchi, C. Castillo, D. Donato, and S. Vigna. Query suggestions using query-flow graphs. In *Proceedings of the 2009 workshop on Web Search Click Data*, WSCD '09, pages 56–63, New York, NY, USA, 2009. ACM.

[5] N. Craswell and M. Szummer. Random walks on the click graph. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, pages 239–246, New York, NY, USA, 2007. ACM.

[6] A. Druin, E. Foss, L. Hatley, E. Golub, M. L. Guha, J. Fails, and H. Hutchinson. How children search the internet with keyword interfaces. In *IDC '09: Proceedings of the 8th International Conference on Interaction Design and Children*, pages 89–96, New York, NY, USA, 2009. ACM.

[7] C. Eickhoff, P. Serdyukov, and A. P. de Vries. Web page classification on child suitability. In *CIKM*, pages 1425–1428, 2010.

[8] W. Gao, C. Niu, J.-Y. Nie, M. Zhou, K.-F. Wong, and H.-W. Hon. Exploiting query logs for cross-lingual query suggestions. *ACM Trans. Inf. Syst.*, 28:6:1–6:33, June 2010.

[9] K. Gyllstrom and M.-F. Moens. Wisdom of the ages: toward delivering the children's web with the link-based agerank algorithm. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, pages 159–168, New York, NY, USA, 2010. ACM.

[10] D. Liu, X.-S. Hua, L. Yang, M. Wang, and H.-J. Zhang. Tag ranking. In *18th International World Wide Web Conference*, pages 351–351, April 2009.

[11] Y. Ma, H. Lin, and Y. Lin. Selecting related terms in query-logs using two-stage simrank. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, CIKM '11, pages 1969–1972, New York, NY, USA, 2011. ACM.

[12] Q. Mei, D. Zhou, and K. Church. Query suggestion using hitting time. In *Proceedings of the 17th ACM conference on Information and knowledge management*, CIKM '08, pages 469–478, New York, NY, USA, 2008. ACM.

[13] I. Szpektor, A. Gionis, and Y. Maarek. Improving recommendation for long-tail queries via templates. In *Proceedings of the 20th international conference on World wide web*, WWW '11, pages 47–56, New York, NY, USA, 2011. ACM.

[14] S. D. Torres, D. Hiemstra, and P. Serdyukov. Query log analysis in the context of information retrieval for children. In *SIGIR*, pages 847–848, 2010.

[15] X. Wang and C. Zhai. Mining term association patterns from search logs for effective query reformulation. In *CIKM '08*, pages 479–488, New York, NY, USA, 2008. ACM.

[16] R. Wetzker, C. Zimmermann, and C. Bauckhage. Analyzing social bookmarking systems: A del.icio.us cookbook. In *Mining Social Data (MSoDa) Workshop Proceedings*, pages 26–30. ECAI 2008, July 2008.

[17] L. Zhang and Y. Zhang. Interactive retrieval based on faceted feedback. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, pages 363–370, New York, NY, USA, 2010. ACM.