http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 6 (4): 584-588, 2007
ISSN 1812-5638

Improving Information Retrieval Precision Using Query Log Mining and Information Scent

Punam Bedi and Suruchi Chawla
Department of Computer Science, University of Delhi, India

Abstract: Effectiveness of the Information Retrieval techniques is becoming a great challenge due to enormous growth of web data. Information Retrieval precision can be improved if the information is retrieved using the information need of the user. This research proposes a method to improve the retrieval precision of search engine by using information scent and multimodal feature of clicked pages in query log mining. The algorithm is based on clustering query sessions using information scent of clicked URLs in the sessions which model the information need associated with the query sessions. In this approach search is driven by the information need of input query. Retrieval precision is improved by boosting the rank of clicked pages in retrieved multimodal web pages for the input query using the measure of similarity of input query to click pages in the selected cluster of query sessions. Performance of the proposed algorithm is evaluated with an experimental study of query log mining of AlltheWeb search engine and it confirms the improvement in information retrieval precision.

Key words: Information scent, information retrieval, multimodal, clustering, search engine, pagerank

INTRODUCTION

Current search tools retrieve too many documents of which only a small fraction is relevant to the user query. Furthermore the most relevant documents do not necessarily appear in top ranking (Gudivada *et al*., 1997). A fundamental task of web search engines is to rank the set of pages returned to the user's query in such a way that most relevant pages to the query appear in the beginning of the search result. Early search engines used keyword of queries and text in web pages to measure the similarity to rank pages. The vector models in (Baeza-Yates and Ribeiro-Neto, 1999) were used for this purpose. However this approach did not achieve good result. Some algorithms make use of links in web to find the quality of page (Kleinberg, 1998; Lawrence *et al*., 1998).

Search engines log keeps track of queries and URLs selected by the users when they are finding useful data through the search engines. Algorithm that use query log is given in (Baeza-Yates *et al*., 2004) where clusters of query sessions is made to identify the group of user preferences of significant size from which useful ranking can be derived. This approach works for those queries that have already been asked, or in other words which are already present in clusters. Secondly while clustering query sessions information need associated to the query sessions is not considered from the point of view of user interest. To overcome these problems the solution proposed in this research is to cluster query sessions using information need modeled by information scent and multimodal features of clicked URLs in the sessions. This will generate the cluster of query sessions with similar information needs. The input query is then used to select the clusters containing user preferences from which useful ranking can be derived using cosine measure of similarity of clicked URLs to the input query in selected cluster of query sessions.

This research aims to improve Information Retrieval precision by adjusting the page rank of retrieved pages by using preferences of user for those query sessions in the log whose associated information need is similar to the information need associated to the input query. In order to cluster query sessions with similar information need, information scent and multimodal vector representation of each clicked page in the query sessions is used.

More unique is the page to the session, more is the information scent associated to it in determining the information need of the session. Another parameter that is taken in calculating the information scent of the clicked pages is the time spent on the clicked pages. Thus both the parameters decide the weightage for the pages in determining the information need associated to query sessions in which they are appearing. K-Means algorithm is used to cluster the query sessions with similar need and each cluster is represented by mean value of vectors contained in a given cluster.

Information scent: On the web, users forage for information by navigating from page to page along the web links. Their actions are guided by their information need. Information scent is the subjective sense of value

Corresponding Author: Punam Bedi, Department of Computer Science, University of Delhi, India

and cost of accessing a page based on perceptual cues with respect to the information need of user. More the page is satisfying the information need of user, more will be the information scent associated to it. The interactions between user need, user action and content of web can be used to infer information need from a pattern of surfing (Chi *et al*., 2001; Pirolli, 1997).

Information scent metric: The Inferring User Need by Information Scent (IUNIS) algorithm (Chi *et al*., 2001) is used to weigh each page vector using the combined effect of two factors according to IUNIS algorithm. The factors are page access TF.IDF weight and TIME that are used to quantify the information scent associated with the page. In TF.IDF the TF term frequency correspond to the access frequency of the page in the given query session and the IDF inverse document frequency correspond to the ratio of total query sessions in the log to the number of query sessions in which this page is accessed. This helps to reduce the weight of those pages that are accessed in many query sessions and may not be very relevant to the information need associated with the current session in which they are appearing. The second factor that is taken is TIME spent on a page in a given query session. By including the time more weightage is given to those pages which consume more user attention.

Multimodal clustering using information scent: Multi modal clustering is a technique which utilizes multiple data feature of the page P_d to produce clusters. The modality vectors are Content, URL, Inlink, Outlink (Heer and Chi, 2002). These individual subvectors are concatenated to form a single multi modal vector.

$P_d = (Content_d, URL_d, Inlink_d, Outlink_d)$ for each document d.

$Content_d$: The content subvector of a page P_d is a weighted keyword vector describing the content of the page P_d.

URL_d: The URL subvector of a Page P_d is URL token keyword vector modeling the URL of the Page P_d.

$Outlink_d$: The Outlink subvector of a page describes which pages are reachable from this page P_d.

$Inlink_d$: The Inlink subvector describes which pages link to this page P_d.

For each modality, we use tf.idf weighing scheme to represent the modality subvector for a given page P_d. The importance of each item in a given page P_d is calculated using TF.IDF (term frequency* inverse document frequency) with respect to the Vocabulary V chosen for a particular modality. Vocabulary V is a set of distinct items found in all distinct clicked pages in whole query log relevant to a particular modality. The composition of Vocabulary set V will be different for each chosen modality. The TF.IDF item weight is calculated as number of times an item appears in the given page weighted by the ratio of the number of all pages to the number of the pages that contain the given item.

The information scent associated with the given page is calculated by using two factor i.e. page access TF.IDF and TIME . Each query session is constructed as linear combination of vector of each page P_d scaled by the weight s_{id} which is the information scent associated with the page P_d in session i. That is

$$Q_i = \sum_{d=1}^{n} s_{id} P_d \quad (1)$$

In above formula n is the number of distinct clicked pages in the query log and s_{id} (information scent) is calculated for each page P_d in a given session i as follow.

TF. IDF (P_{id}): In page access TF.IDF(P_{id}), TF correspond to the page P_d frequency f_{Pid} in a given query session Q_i and IDF correspond to the ratio of total number of query sessions M in the whole log to the number of query sessions m_{Pd} that contain the given page P_d.

Time (P_{id}): The time spent on the given page P_d in a given session Q_i.

In a query session Q_i the s_{id} is calculated as follows

$$s_{id} = TF.IDF(P_{id}) * TIME(P_{id}) \; \forall d \in \{1..n\} \quad (2)$$

$$TF.IDF(P_{id}) = f_{Pid} * \log(M/m_{Pd}) \text{ where } d \in 1..n$$

Vector Model is used for representing each modality of the page P_d in all query sessions. Each page P_d is represented by vector $(w_{1,d}, w_{2,d}, w_{3,d}, \ldots w_{v,d})$ using TF.IDF where v is the number of items in the vocabulary set V where set V will be different for each modality.

Each query session Q^m_i is obtained as weighted vector using formula (1) for each modality m. This vector is modeling the information need associated with the query session Q_i for modality m.

The similarity of query sessions Q_i and Q_j in K-means clustering is calculated as follows.

$$Sim(Q_i, Q_j) = \sum w_m \cos(Q^m_i, Q^m_j) \; \sum w_m = 1 \quad (3)$$

$$m \in M1$$

$$M1=\{ Content, Url, Inlink, Outlink\}$$

Clustering queries: Query sessions are clustered using K-means algorithm because of its good performance for document clustering (Zhao and Karypis, 2002a). Furthermore query sessions in our approach are similar to the vectors of web pages. Thus clusteringqueries in our approach are similar to those for clustering pages.

The quality of resulting clusters is measured by a score or criterion function used by common vector space implementation of K-means algorithm (Zhao and Karypis, 2002b). The function measures the average similarity between vectors and the centroid of clusters that are assigned to. Let C_p be a cluster found in a K-way clustering process ($p \in 1..K$) and let c_p be the centroid of pth cluster .The criterion function I is defined as follows

$$I = 1/M \sum_{p=1}^{K} \sum_{v_i \in C_p} sim(v_i, c_p) \quad (4)$$

where v_i is the vector representing some query session belonging to the cluster C_p and centroid c_p of the cluster C_p is defined as given below.

$$c_p = \left(\sum_{v_i \in C_p} v_i \right) / |C_p|$$

where M is the total number of query sessions in all clusters and $|C_p|$ denotes the number of query sessions in cluster C_p. sim (v_i, c_p) is calculated using cosine measure.

Ranking algorithm for information retrieval: The ranking algorithm is based on clustering process that defines neighborhood of similar query sessions using information need modeled by information scent and multi modal vector of clicked URLs in the query sessions. Each query session consist of a query along with the clicked URLs in its answer.

$$\text{Query session} = (\text{query}, (\text{clicked URLs})^+)$$

where clicked URLs are those URLs which user clicked before submitting another query.

Algorithm

- Offline preprocessing phase at regular and periodical intervals
- Extract the queries and associated clicked URLs from the query log.
- Preprocess the extracted Queries to find the query sessions.
- Model the information need associated to each query session using information scent and weighted vector of multimodal pages in the session

$$Q_i = \sum_{d=1}^{n} s_{id} P_d$$

$$s_{id} = TF.IDF(P_{id}) * TIME(P_{id}) \; \forall d \in \{1..n\}$$

$$TF.IDF(P_{id}) = f_{Pid} * \log(M/m_{Pd}) \text{ where } d \in 1..n$$

- Cluster the query session using information need associated to each query session using formula

$$Sim(Q_i, Q_j) = \sum_{m \in M1} w_m \cos(Q^m_i, Q^m_j) \qquad \sum w_m = 1$$

$$M1 = \{Content, Url, Inlink, Outlink\}$$

- For each cluster C_j create a list U_j of k most frequent URLs using frequency of URLs in cluster C_j.

Online processing phase

- Find the C_j cluster to which input query q belong.
- if no cluster found then
  Use TF.IDF weight for the content modality of input query q. Find the C_j cluster which is most similar to the term weight vector of the input query as per the threshold value set for the similarity measure using content modality .
- If some C_j cluster found then use the URLs in set U_j associated with selected cluster C_j to boost their rank in the retrieved set for the input query q else otherwise do nothing.

The rank of pages in set U_j is calculated using the similarity measure of each page vector u to the input query vector q using content modality only such that those pages with high value of similarity to input query q are ranked higher than those pages with low value of similarity to input query q where sim(u,q) is calculated using cosine measure between vector u and q.

$$NewRank(u) = \{sim(u,q) : \forall u \in U_j\} \quad (5)$$

$$sim(u,q) = \cos(u,q)$$

Experimental study: Experiments were performed on AlltheWeb search engine log containing several thousand queries out of which we have extracted 15095 entries. The query log of AlltheWeb search engine contains the following fields.

- IP Address
- Time of the Day
- Query terms
- Clicked URLs

On submission of the input query, AlltheWeb search engine returns a result page consists of relevant URLs with information about URLs. The URLs are ranked in order of relevance to the input query as determined by the AlltheWeb internal relevance function. In the experiments only those queries in the query log are selected which have at least one click in their answers. Query sessions considered consist of query terms along with the clicked URLs. The clicked URLs are those URLs which user clicked before he submits another query. The number of distinct URLs in the query log was found to be 12025. The query log is preprocessed to get 2586 query sessions. In this experiment we have used the content modality of each clicked page in the query sessions for simplicity reason however other feature can be added later on.

$$Sim(Q_i,Q_j)= \cos(Q^{content}{}_i,Q^{content}{}_j)$$

$$w^{content}=1, w^{url}=0, w^{inlink}=0, w^{outlink}=0$$

The query sessions were clustered using K-Means algorithm. The K-Means algorithm was executed several times for different values of K and criterion function was computed for each value of K. The criterion function was found to have maximum value at K = 85. The similarity of vectors is measured using cosine formula for weighted term vector. The similarity threshold value was set to 0.5. The experiments were performed on queries selected from three domain mainly entertainment, academics and sports. The sample of queries taken in each of these domains from query log is given below in Table 1.

The experiment is performed on Pentium IV PC with 512 MB RAM under Windows XP using C++ software. The K-Means Algorithm is executed to generate cluster of query sessions. Each cluster of query sessions is represented by mean value of vector of terms.

Some of the queries in cluster to which media player belong or similar to it is given in Table 2.

Table 1: Sample of queries in selected domains

Domain	Queries used in
Entertainment	Free pics, online audio stores
	Free download mp3, skies of arcadia pictures
	vcd files, mpeg movies, free mpeg
	chatrooms
	dvd pc software
	yahoocard greeting, greeting cards
Sports	Grand American road racing series
	Arena football, South dakota wrestling
	Major league baseball tryouts, kit car
	Arena football, produnova
	Wnrb
	Panoz racing, lowrider bicycles
Academics	Cgi perl tutorial, sql tutorial, tutorial oracle
	Windows 2000 tutorial, macros, templates
	Weblogs, solar system, perl tutorial
	Relative hnmidity table, macromedia director
	Xor gate, flash online
	Education cartoons

Table 2. Queries in cluster to which Media player belongs

Query	Other queries in cluster
Media player	Microsoft media player 7
	Windows media player
	Free download mp3
	Media player codec
	Porn free windows media player
	Media player nt
	Windows media audio, Xing mediaplayer
	Windows media video macintosh codec
	Freeware mpeg2 player, ms mediaplayer

Performance of ranking algorithm: The performance of ranking algorithm based on query log mining using information scent is evaluated by anonymous users having knowledge in domain from which queries are selected. The performance is evaluated on trained and untrained set of queries belonging to each of the domains considered. The trained set of queries are those queries which are present in query sessions cluster and untrained set of queries are those queries which are not present in query log. The experiment is performed on both trained and untrained set of queries separately. The performance is evaluated by using average precision of queries for each domain. The experiment is performed using 28 trained queries and 32 untrained queries. The average precision is calculated for first top 10 result set and users mark the relevant documents within the list of top 10 URLs retrieved for a given query using both base algorithm approach and proposed algorithm approach.

The Fig. 1 and 2 showing the average precision of both base algorithm and proposed algorithm for both trained and untrained set of queries is given below. The above experiment shows that information retrieval precision is improved for both trained set of queries and untrained set of queries using proposed algorithm. The average precision of untrained set of queries shows significant improvement and looking forward to improve the precision further in future.

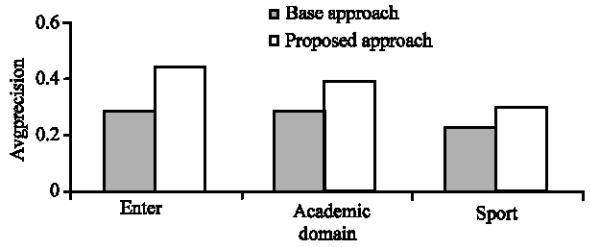

Fig. 1: The average precision of base and proposed algo on trained queries

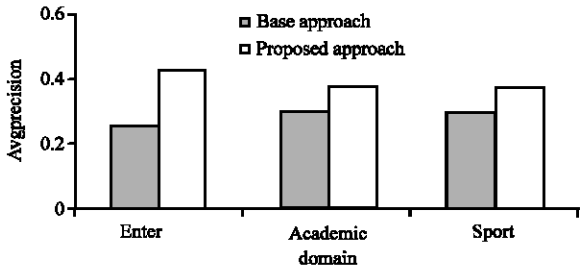

Fig. 2: The average precision of base and proposed algo on untrained queries

CONCLUSION

In this study efforts have been made for improving the information retrieval precision by modeling the information need associated to the query sessions using information scent. An attempt is made to improve the retrieval precision by boosting the rank of retrieved pages by determining the information need associated to the query and use the users preferences with similar need to boost the rank of retrieved pages. Information scent helps to derive the information need associated to the user request and enables to use preferences of users with similar needs to satisfy current user need efficiently. Experimental results confirm the improvement of the precision of information retrieval using proposed algorithm.

ACKNOWLEDGMENT

We extend special thanks to Amanda Spink and Bernard J. Jansen for providing us AlltheWeb search engine query log without which we could not have conducted this research.

REFERENCES

Baeza-Yates, R. and B. Ribeiro-Neto, 1999. Modern Information Retrieval, Addison Wesley.

Baeza-Yates, R., C.A. Hurtado and M. Mendoza, 2004. Query clustering for boosting web page ranking . In: Advances in Web Intelligence, Second International Atlantic Web Intelligence Conference, AWIC 2004, pp: 164-175.

Chi, E.H., P. Pirolli, K. Chen and J. Pitkow, 2001, Using Information Scent to model User Information Needs and Actions on the Web. In: Proc. ACM CHI 2001 Conference on Human Factors in Computing Systems, pp: 490-497.

Gudivada, V.N., V.V. Raghavan, W.I. Grosky and R. KasanaGottu, 1997. Information Retrieval on World Wide Web. IEEE. Expert.

Heer, J. and E.H. Chi, 2002. Separating the Swarm: Categorization method for User Access Session on the Web. In: Proc. ACM CHI 2002 Conference on Human Factor in Computing System, pp: 243-250.

Kleinberg, J., 1998. Authorative sources in a hyper linked environment. ACM-SIAM Symposium on Discrete Algorithm (SODA), 1998.

Lawrence, P., S. Brin, R. Motwani and T. Winograd, 1998. The page rank citation ranking: Bringing order to the web. Stanford Digital Library Technologies Project.

Pirolli, P., 1997. Computational models of information scent-following in a very large browsable text collection. In: Proc. ACM CHI 97 Conference on Human Factors in Computing Systems, pp: 3-10.

Zhao, Y. and G. Karypis, 2002a. Comparison of agglomerative and partitional document clustering algorithms. In: SIAM Workshop on Clustering High-dimensional Data and its Applications.

Zhao, Y. and G. Karypis, 2002b, Criterion functions for document clustering. Technical report, 2002. University of Minnesota, Minneapolis, MN.