Topic-Driven Web Search Result Organization by Leveraging Wikipedia Semantic Knowledge

Xianpei Han Jun Zhao
National Laboratory of Pattern Recognition, Institute of Automation
Chinese Academy of Sciences, Beijing, China
+86 10 8261 4468
xianpei@nfs.iscas.ac.cn jzhao@nlpr.ia.ac.cn

ABSTRACT

Effective organization of web search results can greatly improve the utility of search engine and enhance the quality of search results. However, the organization of search results is difficult because the sub-topics of a query are usually not explicitly given. In this paper, we propose a novel topic-driven search result organization method, which can first detect the sub-topics of a query by finding the coherent Wikipedia concept groups from its search results; then organize these results using a topic-driven clustering algorithm; in the end we score and rank the topics using the support vector regression model. Empirical results show that our method can achieve competitive performance.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Clustering, Search process

General Terms: Algorithms, Experimentation

Keywords

Search Result Organization, Topic Detection, Topic Ranking, Topic-Driven Clustering.

1. INTRODUCTION

With the explosion of Web, search engine has been the critical way for finding the relevant information from billions of web pages. Given a user query, most today's search engines present the relevant documents as a flat ranked list. The ranked list presentation, however, although works well for the simple and navigational queries, can often fail in addressing vague, broad or ambiguous queries, where the search result is often the mixture of web pages belonging to different subtopics of the given query. For example, as shown in Table 1, when we query Google with "Jaguar" and "Information Retrieval", both their returned results contain more than 10 topics.

Query	Subtopics
Jaguar	*Jaguar(animal cat), Jaguar(car), Atari Jaguar, Jaguar(film), Jaguar(Mac OS X),…*
Information Retrieval	*Book, Algorithm, Conference, Paper, Software, Ranking, Crawling, Clustering, ….*

Table 1. The sub-topics of two selected queries

In order to improve the utility of search engine, a possible solution is search result organization, which groups search results into clusters according to the subtopics of a given query, so that a user can easily navigate into a particular interesting subtopic. Furthermore, as shown in Hearst and Pedersen [1], search results organization can also enhance the quality of search results, since relevant documents tend to be more similar to each other.

The search result organization, however, is difficult because the subtopic knowledge of a query is usually not explicitly given. Conventionally, the traditional methods organize search results based on only content similarities ([1][5]) or whether some salient phrases are shared ([6][7][8][9]). By taking no topic knowledge (sub-topics of a query) into consideration, the traditional clustering-based methods often generate clusters which do not correspond to semantically meaningful topics.

In order to resolve the traditional methods' deficiencies, this paper proposes a novel topic-driven search result organization method. The start point of our method is that a topic can be represented efficiently using a set of coherent Wikipedia concepts. Starting from this point, given the query and the ranked list of documents (typically a list of titles and snippets) returned by a certain web search engine, our method works as follows: Firstly, the sub-topics of the query are detected by finding the coherent Wikipedia concept groups through a community detection process. Secondly, the search results are clustered according to the detected topics using a topic-driven clustering algorithm. Finally, in order to present the topics in a user-friendly way, the topics are scored and ranked by combining their properties (including *Salience*, *Predictiveness*, *Coherence* and *Distinguishness*) using the support vector regression model.

This paper is organized as follows. The related work is reviewed in Section 2. The problem is defined in Section 3. In Section 4, we introduce our topic detection method. The topic-driven search result organization method is described in Section 5. The experimental results are presented and discussed in Section 6.

CIKM'10, October 26-30, 2010, Toronto, Ontario, Canada.
Copyright 2010 ACM 978-1-4503-0099-5/10/10...$10.00.

Finally we conclude the paper and give some future work in Section 7.

2. RELATED WORK

The problem of search result organization has been investigated in a number of previous researches. According to how well they generate user-friendly labels of clusters, the traditional methods can be classified into two categories: the data-centric method and the label-centric method. The data-centric methods ([1][5]) usually group search results into clusters using conventional document clustering algorithms, then produce some kind of textual label of resulting clusters for end users. Unlike the data-centric methods, the label-centric methods put their emphasis on the quality of cluster labels. One of the pioneer researches was the Suffix Tree Clustering (STC) method proposed by Zamir and Etzioni [6], which selected the frequent phrases as the labels of clusters. The STC method was extended in a follow-up algorithm called HSTC [7]. Another method was the Lingo method proposed by Osiriski [8], which found the cluster labels through the singular value decomposition process. Lawrie and Croft [9] also proposed a method which selected cluster labels based on two specific document statistics named topicality and predictiveness.

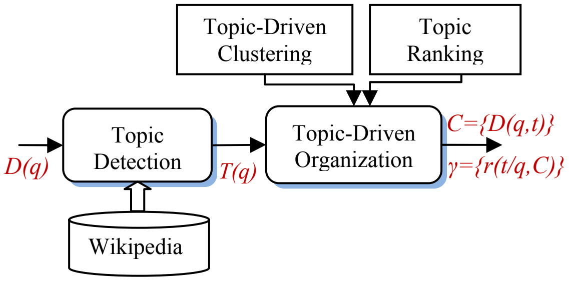

Figure 1. The proposed method's framework

3. PROBLEM FORMALIZATION

In this section, we formulize the topic-driven search result organization problem. Conventionally, as shown in Figure 1, a topic-driven search result organization system is defined as a six-tuple $O=\{q,D,T,C,\delta,\gamma\}$, where:

q is the query, such as "Jaguar" and "Information Retrieval";

$D(q)=\{d_1, d_2, \dots, d_N\}$ is the top N search results of query q, returned by a certain search engine such as Google and Yahoo!;

$T(q)=\{t_1, t_2, \dots, t_k\}$ is the subtopics of query q;

$C=\{D(q,t)\}$ is the clusters of search results, each cluster $D(q,t)=\{d_{t1}, d_{t2}, \dots ,d_{tm}\}$ is a subset of the search results, and corresponds to a specific topic t within $T(q)$;

$\delta : D(q)\times T(q)\rightarrow C$ is the clustering algorithm, which groups search results into clusters based on the sub-topics of a query;

$\gamma=\{r(t\,|\,q,C)\}$ is the topic ranking algorithm, which scores the importance and relevance of each topic.

Based on the above formalization, the main task of our topic-driven search result organization is to design the clustering algorithm $\delta : D(q)\times T(q)\rightarrow C$ and the topic ranking algorithm $\gamma=\{r(t\,|\,q,C)\}$. On the other hand, because the topic knowledge $T(q)$ is not given, an additional topic detection step is also needed.

4. TOPIC DETECTION

In this section, we demonstrate how to detect the topics within search results by leveraging Wikipedia semantic knowledge. In this paper a topic is represented as a set of coherent Wikipedia concepts, e.g., the *Jaguar(animal cat)* topic can be represented as *{Animal, Jaguar, Big Cat, Leopard, Wildcat, Felis, …}*. Based on this idea, the topic detection is regarded as a task of finding the coherent Wikipedia concept groups within the search results, which is composed of three steps: (1) Wikipedia concept extraction; (2) semantic graph building; (3) community detection. In the following we respectively describe each of the three steps.

Wikipedia Concept Extraction. The goal of this step is to extract Wikipedia concepts from search results. In this paper, we extract Wikipedia concepts using the method described in Medelyan et al. [2].

Semantic Graph Building. In this step we model the semantic relations between the Wikipedia concepts as a semantic graph, which is defined as follows:

> ***Semantic graph*** *is an un-weighted graph where each vertex is a Wikipedia concept, each edge between a pair of vertices means that the two Wikipedia concepts corresponding to these vertices are semantically related.*

To build the semantic graph, we first measure the semantic relatedness *sr* between Wikipedia concepts using the method proposed in Witten and Milne [4], then use a semantic relatedness threshold *ST* to determine whether two concepts are semantically related, i.e., two Wikipedia concepts are considered semantically related if the semantic relatedness between them is larger than ST. The value of ST is set to 0.3 in this paper through a learning process. For demonstration, a semantic graph of four Jaguar search results is shown in Figure 2.

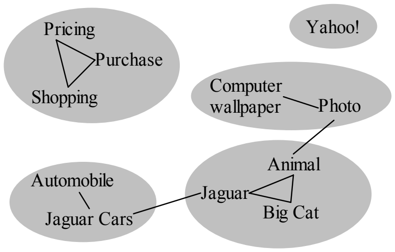

Figure 2. A semantic graph and its community detection results

Community Detection. Given the semantic graph, the goal of this step is to find the coherent groups of Wikipedia concepts, with each resulting group corresponding to one specific topic.

We observe that the semantic relations are dense between the Wikipedia concepts within the same topic; while the semantic relations are sparse between the Wikipedia concepts across different topics. Based on the above observation, we can regard the topics as the community structures in networks [3]. Therefore we can detect the topics by finding the communities in the semantic graph. We employ the algorithm described in

Newman and Girvan [3] to discover the communities in a semantic graph. For demonstration, we show the community detection result of the semantic graph in Figure 2.

5. TOPIC-DRIVEN SEARCH RESULT ORGANIZATION

In the above section, we have detected a set of topics *T(q) = {t₁, t₂, ..., t_k}* for the query *q*, with each topic *t* represented as a set of coherent Wikipedia concepts *{c₁, c₂, ..., c_m}*. In this section, we organize search results according to the detected topics *T(q)* through the following two steps described in the following subsections:

1) We group search results into clusters through a topic-driven search result clustering algorithm;

2) We rank topics using the support vector regression model, so that the relevant and important topics will be presented at the top positions.

5.1 Topic-Driven Search Result Clustering

Our topic-driven search result clustering algorithm groups search results into clusters by assigning documents to topics based on their content similarity. The documents assigned to the same topic *t* is viewed as a single cluster, denoted as *D(q,t)*, or *D(t)* for simplicity.

The first step is to compute the similarity between a document *d* and a topic *t*. We first represent both the document and the topic as a weighted Wikipedia concept set *{(c₁, w(c₁)), (c₂, w(c₂)), ..., (c_m, w(c_m))}*, where *w(c)* is the weight of concept *c*. For a topic *t*, the concepts are weighted using the following formula:

$$w(c) = |t|^{-1} \Big(\sum_{c_i \in t, c_i \neq c} sr(c, c_i) \Big)$$

which means the weight of a concept is its average semantic relatedness to all the other concepts within the same topic. For a document *d*, we first obtain its Wikipedia concept set representation by extracting all the Wikipedia concepts within its content (title and snippet) using the same method described in Section 4.1, then weight the concepts using the above formula.

Based on the weighted Wikipedia concept set representation, we compute the similarity between a topic *t* and a document *d* as:

$$SIM(t,d) = \sum_{c_i \in t} \sum_{c_j \in d} w(c_i) w(c_j) sr(c_i, c_j) \Big/ \sum_{c_i \in t} \sum_{c_j \in d} w(c_i) w(c_j)$$

Based on the computed similarities, we assign a document *d* to a topic *t* if their similarity *SIM(t,d)* is larger than a similarity threshold *CT*(a document may be assigned to multiple topics). In this paper the value of *CT* is set to *0.26* through a learning process.

Finally, we generate the label for each cluster *D(t)*. In our method, we simply choose the Wikipedia concept with the maximal weight in topic *t* as the *D(t)*'s label.

5.2 Topic Ranking

Till now, we have organized the search results into clusters according to the detected topics. However, not all the topics are equally relevant and important to users: some topics may not be relevant to the query; some topics may not be salient in the search results, etc. So a topic ranking step is critical for presenting the relevant and important topics at the top positions to users. In this section, we propose an algorithm which can rank topics by giving each topic an importance score. The topic ranking algorithm first measures four properties of a topic (including *Salience*, *Predictiveness*, *Coherence* and *Distinguishness*) which are supposed to be the important factors for topic ranking; then combines all these properties using the support vector regression model. The detail description is shown in follows.

Salience. The salience measures how salient a topic is in the search results. Intuitively, a topic *t* is salient if it is the main topic of the documents within *D(t)* and contained in many search results. So we can measure the salience of a topic using the following two measures:

The Average Concept Frequency in Topic Cluster:

$$ACF(t) = \sum_{d \in D(t)} CF(t,d) \Big/ |D(t)|, \text{ where } CF(t,d) = |t \cap d| / |d|$$

The Document Frequency of Topic *t*:

$$DF(t) = |D(t)| / |D(q)|$$

Coherence. The coherence measures the quality of a topic, that is, how coherent a topic is. Intuitively, a topic *t* is coherent if the concepts within it are highly semantically related and the documents assigned to this topic are highly similar to each other, which are measured as two individual values:

The Concept Coherence:

$$T\_Cohen(t) = \sum_{c_i, c_j \in t} sr(c_i, c_j) \Big/ |t|^2$$

The Cluster Coherence:

$$C\_Cohen(t) = \sum_{d_i, d_j \in D(t)} SIM(d_i, d_j) \Big/ |D(t)|^2$$

Predictiveness. The predictiveness measures how well a user can deduce the content of the cluster *D(t)* by glancing through the representation of topic *t*, which is computed as the average similarity between the topic *t* and the documents assigned to it:

$$Pred(t) = \sum_{d \in D(t)} SIM(t,d) \Big/ |D(t)|$$

Distinguishness. The distinguishness measures how well a topic can distinguish the documents assigned to it from the documents not assigned to it, which is calculated as the average similarity between the topic *t* and the documents not assigned to it:

$$Dis(t) = \sum_{d \notin D(t)} SIM(t,d) \Big/ |D(q) - D(t)|$$

Combining all the Evidences by SVM Regression. Given the four properties of a topic, in the following we combine them and calculate a single importance score for each topic. In this paper, we use support vector regression ([10]) to combine the four properties, which can balance the four properties' values and output a final importance score. Finally we rank the topics in *T(q)* according to their importance scores.

6. EXPERIMENTS AND DISCUSSIONS

In the following, we first explain the general experimental settings in Section 6.1, then evaluate and discuss the experiment results in Section 6.2.

6.1 Experimental Settings

Wikipedia We use the English Wikipedia version released on Mar. 6, 2009, which contains more than 6,600,000 distinct concepts.

Dataset We adopt the AMBIENT [7] dataset for evaluation, which consists of 44 ambiguous and broad queries. For each query, the AMBIENT dataset provide a set of subtopics (totally 790 subtopics) and a list of 100 ranked documents returned by a search engine, with all the documents annotated with the subtopic information.

We also build a dataset for training and evaluating our topic ranking algorithm: We manually annotate the topics detected by our method with three importance scores: *1.0*, *0.5* and *0*, where the score *1.0* means that this topic is very important; *0.5* means that this topic is relevant to the query but not very important; *0* means that this topic is not relevant to the query.

Evaluation Criteria To compare the clustering performance of our method with other search result clustering methods, we adopt the same method as in [1], that is, we compare the quality of the best cluster, which is defined as the one with the largest number of relevant documents. We use Precision at top 5 documents (**P@5**) in the best cluster as the primary measure to compare different methods. Except for the **P@5**, we also provided the **P@10** and the **Mean Reciprocal Rank (MRR)** as additional precision metrics. The **Recall** of the best cluster is also provided. To evaluate the performance of the topic ranking algorithm, we adopt the precision at the top N topics (**TP@N**).

6.2 Experiment Results

We compare our method with three baselines: 1) The original ranked list of search results returned by a search engine, where all the search results are viewed as a single cluster – we denoted it as **Ranked_List**; 2) The Suffix Tree Clustering method proposed by Zamir and Etzioni[6] – we denoted it as **STC**; 3) the LINGO method described in Osiriski et al. [8] – we denoted it as **LINGO**.

	P@5	P@10	MRR	Recall
Ranked_List	0.11	0.10	0.04	1.00
STC	0.45	0.37	0.16	0.85
LINGO	0.48	0.30	0.29	0.70
Our Method	**0.53**	**0.38**	**0.36**	**0.92**

Table 2. Performance results of baselines and our method

Kernel Type	TP@All	TP@20	TP@10	TP@5
Linear	0.05	0.26	0.49	**0.71**
Polynomial	0.05	0.22	0.42	**0.67**
RBF	0.05	0.24	0.46	**0.75**

Table 3. Performance results of Topic Ranking

6.2.1 Overall Results

We compare our method with all the three baselines. The overall performance is shown in Table 2. From the performance results in Table 2, we can see that:

1) The search result organization can greatly improve the performance and utility of search engine: compared with the Ranked_List baseline, all the three search result organization methods obtained significant performance improvements: the STC, LINGO and Our Method achieve respectively 34%, 37%, 42% P@5 improvement.

2) Our method can effectively organize the search results. Compared with the STC baseline, our method gets 8% P@5 improvement; compared with the LINGO baseline, our method gets 5% P@5 improvement.

3) Our method worked well in grouping search results into clusters which are not only precise but also with high-recall as well. Compared with the STC and the LINGO, our method can obtain not only high precision but also high recall as well: the MRR is 0.36, 7% improvement over the LINGO; the Recall is 0.92, 7% improvement over the STC.

To evaluate the efficiency of topic ranking, we show its performance in Table 3 (we use the topics of 30 queries for training, 14 queries for testing). From the results in Table 3, we can see that:

1) The topic ranking is critical for search result organization: only 5% topics are relevant to its query (shown as the TP@ALL). Without topic ranking, a user will be hard to choose the relevant and important topics.

2) Our topic ranking algorithm is effective. It can achieve high precision at the top positions: 0.75 precision at the top 5 topics, 0.49 at the top 10 topics.

7. CONCLUSIONS

In this paper we propose a novel topic-driven search result organization method, which firstly detect the topic knowledge of a query from its search results, then organize the search results according to the detected topics. By detecting and integrating the topic knowledge, our method can achieve competitive results.

For future work, we will organize the search results through detecting a hierarchical topic structure, which can further improve the utility of organized search results.

8. Acknowledgments

The work is supported by the National Natural Science Foundation of China under Grants no. 60875041 and 60673042, and the National High Technology Development 863 Program of China under Grants no. 2006AA01Z144.

9. REFERENCES

[1] Hearst, M. A. and Pedersen, J. O. Reexamining the cluster hypothesis: scatter/gather on retrieval results. In Proc. of SIGIR, 1996.

[2] Medelyan, O., Witten, I. H. and Milne, D. Topic indexing with Wikipedia. In Proc. of the AAAI WikiAI, 2008.

[3] Newman, M. and Girvan, M. Finding and evaluating community structure in networks. Physical review E, vol. 69, no. 2, p. 26113, 2004.

[4] Witten, D. M. and Milne, D. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links, In Proc. of AAAI WikiAI , 2008.

[5] Everitt, B.S. Landau S., and Lesse M. Cluster Analysis, 4th Ed. Oxford University Press, 2001.

[6] Zamir, O. and Etzioni, O. Web document clustering: A feasibility demonstration. In Proc. of SIGIR, 1998.

[7] Masowska, I. Phrase-Based Hierarchical Clustering of Web Search Results. Advances in Information Retrieval, 2003.

[8] Osiriski, S., Stefanowski, J. and Weiss, D. Lingo: Search results clustering algorithm based on singular value decomposition. In Proc.of the IIS: IIPWM'04 , 2004.

[9] Lawrie, D. J. and Croft, W. B. Generating Hierarchical Summaries for Web Searches. In Proc. of SIGIR, 2003.

[10] Smola, A. J. & Schlkopf, B. A tutorial on support vector regression. Statistics and Computing, vol. 14, no. 3, 2004.