Summarization and Expansion of Search Facets*

Aparna Nurani Venkitasubramanian Marie-Francine Moens
Department of Computer Science
Katholieke Universiteit Leuven
Leuven, Belgium
{aparna.nuranivenkitasubramanian,sien.moens}@cs.kuleuven.be

ABSTRACT

We present a novel method for summarization and expansion of search facets. To dynamically extract key facets, the ranked list of search results generated from a keyword search is coupled with the spatial distribution of relevant documents in a hierarchical taxonomy of subject classes. An evaluation of the method based on the relevance and diversity of the produced facets indicates its effectiveness for both summarization and expansion.

Keywords

Selection of search facets, Expansion, Summarization

1. INTRODUCTION

The combination of a 'keyword' and a 'faceted' search has the potential to enhance user experience by providing a better arrangement of search results and aiding further search exploration. However, such a framework poses two key problems: 1) a given query may cover several facets, requiring an aggregation or summarization of the most relevant ones; and 2) a query may cover too few facets necessitating an expansion to include additional facets. We exploit the spatial distribution of topics relevant to a query in a hierarchy together with the relevance ranking of the documents for the query, in order to select search facets that optimize diversity and relevance.

2. SELECTING SEARCH FACETS

We assume that the search results of a query are annotated with subject classes (here *facets* or *nodes*) obtained from a hierarchical taxonomy. In the experiments below, the DMOZ* hierarchy is used. For each query, we define: a set of *activated nodes* that have documents relevant to the query and a set of *presentation nodes* that will be presented to the user as facets relevant for the query.

When the user presents a query, the DMOZ facets associated with the result of the query are first extracted, i.e., the *activated nodes* are identified. Next, if the number of *activated* facets associated with the query is larger than k†, the set of *activated* nodes or facets is summarized by picking the best k candidates. If the number of *activated* facets is less than k, then the set is expanded by adding related facets. The summarization and expansion are carried out using the 'Subtree density' model (Section 3) which takes as input a set of *activated* nodes and produces the *presentation* nodes. For some queries, DMOZ activates not only the lowest level facets, but also some of their ancestors. In such a case to ensure presentation of as many distinct facets as possible, the summarization uses only the descendants, while the expansion uses only the ancestors.

*Full paper published at CORIA 2013 [3]
*http://www.dmoz.org
†k is chosen based on the size of the interface medium and the cognitive load acceptable for a user

DIR 2013, April 26, 2013, Delft, The Netherlands.
Copyright remains with the authors and/or original copyright holders.

3. SUBTREE DENSITY MODEL

This model finds nodes which represent dense clusters of facets, each having many search results important for the query. First, the subtrees associated with the relevant set of activated nodes are extracted. The subtree $\mathbf{S}$ for a node v comprises the node and its descendants (children, grand children etc. until the last level).

Then, one possible candidate to represent a subtree is the medoid identified as the node with the minimum average distance to all the other nodes of the subtree. The distances between nodes in the subtree are computed using a distance metric that captures semantic distances between topics in a hierarchy. Since the basic relations in the taxonomy are the parent-child relations, distance between any two nodes is represented using the connection weights between the parent-child pairs associated. In taxonomy $\mathbf{T}$ with root at level 0, the connection weight D between node v_i at level l and its child v_j at level $l+1$ is as follows:

$$D(v_i, v_j) = 2^{-l} \tag{1}$$

Using this metric, the distance between any two nodes v_m and v_n in $\mathbf{T}$ is defined as the sum of connection weights between all nodes v_x spanning the path between v_m and v_n.

Once the medoids of the subtree have been identified, we must rank them to identify the best k medoids that will be presented. This is done using a score computed in Eq. 2

$$score(m) = \frac{density(\mathbf{S})}{distance(m, \mathbf{S})} \tag{2}$$

where $density(\mathbf{S})$ is given by

$$density(\mathbf{S}) = \frac{\sum_{v \in S} importance(v, \mathbf{R})}{|\mathbf{S}|} \tag{3}$$

where $|\mathbf{S}|$ is the size of the subtree in terms of number of nodes $v \in \mathbf{S}$ and $importance(v, \mathbf{R})$ is computed using the Discounted Cumulative Gain (DCG) [2] over the retrieved Web pages assigned to facet v.

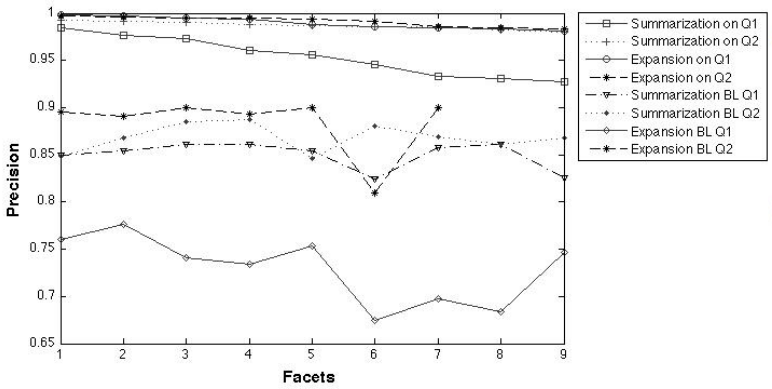

Figure 1: Precision of the summarization and expansion for the nine highest ranked facets for query sets Q1 and Q2

(a)

	Q1	Q2
Queries	1200	1200
Queries with agreement > 80%	1004	995
Queries used for evaluation of summarization	508	523
Queries used for evaluation of expansion	496	472

(b)

	Q1	Q2	BL Q1	BL Q2
Rank1	79.86	94.84	63.80	90.93
Rank2	4.07	2.89	2.04	1.65
Rank3	11.99	1.65	23.98	4.12
Rank4	2.04	0.41	7.92	2.06
Rank5	2.04	0.21	2.26	1.24
Total	100	100	100	100

Table 1: (a) Statistics of the query sets (b) Diversity of facets produced by summarization (% of facet clusters at rank 1..5)

$$importance(v, \mathbf{R}) = rel_1 + \sum_{i=rank(d), i>1, d \in \mathbf{R}} \frac{rel_i}{log_2(i)} \quad (4)$$

where $\mathbf{R}$ is a ranked list of documents retrieved for the query obtained from a search engine, i is the position of the retrieved document in the list, and $rel_i = 1$ if the ith document belongs to facet v and 0 otherwise.

The idea of this score is as follows:

- A node that has lesser distance from every other node of the subtree is a better representative of the subtree;
- A subtree that has a higher density is an important one for the query.

4. EXPERIMENTS AND RESULTS

Two sets of queries have been used for evaluation. The first query set **Q1** contains titles of English Wikipedia articles. The second query set **Q2** comprises real user queries collected by Torres et al. [1]. The queries were submitted to the Bing search engine, restricting the search results to the Web pages from the DMOZ Kids and Teens subdirectory. The subtree density model has been benchmarked against two baseline (BL) models, one for summarization and the other for expansion. The baseline model for summarization uses the top k distinct activated nodes from the ranked results from a search engine, while the baseline model for expansion uses the siblings of the activated nodes for presentation.

The evaluation is based on two aspects- relevance and diversity. First, the facets selected by the model for each query of the two query sets were presented to five Crowdflower[‡] evaluators, who were asked to judge whether the facets produced were relevant to the query. Next, to evaluate diversity of the summarization, we put together two clusters of related facets (that were judged relevant by Crowdflower evaluators)- one for each summarization model, per query for the queries in **Q1** and **Q2**. Then, Crowdflower evalutors were asked to rank these clusters on a scale of 1 to 5 based on the diversity of the facets in the clusters, with rank 1 corresponding to 'Very diverse'. For both relevance and diversity evaluations, only queries for which the agreement among Crowdflower evaluators was over 80% (as reported by Crowdflower) were retained.

The number of queries used for evaluation, the precision and diversity of the model have been indicated in Table 1 and Figure 1. From Figure 1, it is evident that the subtree density model performs better than the baselines in terms of precision (measured by relevance), both in summarization and expansion. Table 1b indicates that the subtree density model also outperforms the baseline (based on ranked results) in terms of diversity. These results are explained by the fact that in our model, important facets come from dense clusters of search results in the taxonomy.

[‡]http://crowdflower.com/

5. CONCLUSIONS

In this paper we have presented the subtree density model for summarizing and expanding search results mapped to a subject taxonomy. Evaluation of the method using human evaluators indicates that it is effective as it optimizes both relevance and diversity. A next step in our research is to develop navigation models for interactive browsing consisting of the presented facets and their corresponding Web pages.

Acknowledgements This research was funded by the Puppy IR project EU FP7 231507.

6. REFERENCES

[1] S. Duarte Torres, D. Hiemstra, and P. Serdyukov. An analysis of queries intended to search information for children. In *Third Symposium on Information Interaction in Context*. ACM, 2010.

[2] K. Järvelin and J. Kekäläinen. IR evaluation methods for retrieving highly relevant documents. In *SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2000.

[3] A. Nurani Venkitasubramanian and M.-F. Moens. Selection of search facets. In *CORIA 2013 - 10th Francophone Information Retrieval Conference*, 2013.