Selection of Search Facets

Aparna Nurani Venkitasubramanian — Marie-Francine Moens

Katholieke Universiteit Leuven
Department of Computer Science
Celestijnenlaan 200A, Heverlee, Belgium
aparna.nuranivenkitasubramanian@cs.kuleuven.be

ABSTRACT. *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 2) a query may cover too few facets necessitating an expansion to include additional facets. In this paper, we propose several novel methods for summarization and expansion of facets. Using the ranked list of search results generated from a keyword search, coupled with the spatial distribution of relevant documents in a hierarchical taxonomy of subject classes, we dynamically extract key facets. An evaluation of the different methods based on the relevance and diversity of the facets indicates that the Subtree density model performs best for both summarization and expansion.*

RÉSUMÉ. *Les recherches par mots clés sur le Web donnent souvent une énorme quantité de pages Web pertinentes. Un cadre qui intègre les avantages á la fois des « mot-clé » et des « facettes » des recherches a des larges avantages pour les utilisateurs Web, car il offre une meilleure organisation des résultats de la recherche et une plate-forme utile pour guider les utilisateurs á trouver les informations pertinentes. Dans un cadre pareil, deux problèmes principaux existent : 1) une requête peut entamer plusieurs facettes, ce qui nécessite une agrégation ou un résumé des facettes les plus pertinentes ; 2) une requête peut couvrir trop peu de facettes nécessitant une recommandation ou une expansion. Dans cet article, nous proposons plusieurs nouvelles méthodes de synthèse et de l'expansion de facettes à partir d'une recherche par mot clé, associées á la distribution spatiale des documents pertinents dans une taxonomie hiérarchique des classes de sujets. Une évaluation des différents modèles basés sur la pertinence et la diversité des facettes indique que le modèle de « subtree density » donne les meilleures résultats.*

KEYWORDS: *Faceted search, Summarization, Expansion*

MOTS-CLÉS : *recherche par facettes*

1. Introduction

Keyword searches result in large answer lists which are difficult to navigate and process for naïve users. Consequently, the problem of presenting and aggregating search results [1] [2] has been a topic of interest [ZAM 99][SUS 08][DOU 11][LI 10]. Towards this goal, one approach is to exploit a hierarchy of facets to better structure the search results. Chen et al. [CHE 00] have proposed a technique to automatically categorize search results using SVMs on the terms most predictive of the facet. Ouamer and Hammache [AHM 10] have devised a model based on a domain-specific ontology with semantic links between documents in the context of e-learning. Agrawal et al. [AGR 09] have implemented an algorithm that re-orders the search results based on the probability of the document and query belonging to a facet as well as the value of the search result for the query given the ranking and its facet. [YEE 03] developed an interface using sentential and phrasal descriptions of images manually organized into hierarchical facets. In contrast to these works, we exploit two parameters 1) the spatial distribution of topics relevant to a query in a hierarchy and 2) the relevance ranking of the documents for the query. The contribution of this paper is twofold: 1) Several novel models that aim at optimizing diversity and relevance of search facets and 2) a novel metric of distance between topics in a hierarchy.

The rest of this paper is organized as follows: Section 2 introduces the preliminaries. Section 3 describes the selection of facets to be presented to the user using options like summarizing or expansion to focus or broaden the search. Section 4 discusses experiments and results. Finally, section 5 presents the future work and conclusions.

2. Preliminaries

Our methods assume that the search results of a query are annotated with subject classes (denoted by *facets*, *nodes* or *facet nodes* in this paper) obtained from a hierarchical taxonomy composed of thousands of classes. In the experiments below the DMOZ [3] hierarchy is used. For each query, we define:

– A set of *activated nodes*, a set of nodes that have documents relevant to the query.

– A set of *presentation nodes*, that will be chosen by some algorithm discussed below, and presented to the user as facets relevant for the query.

2.1. *Estimating the importance of a facet node for a query*

To estimate the $importance$ of a facet node v, the Discounted Cumulative Gain (DCG) [JÄR 00] is computed over the retrieved Web pages assigned to facet v. This score is naturally dependent on the query and its ranked results **R** obtained from a search engine, and is given by Eq. 1:

1. Google's Universal Search http://www.google.com/enterprise/search/solutions_productivity.html
2. Ask's X http://about.ask.com/en/docs/about/askx.shtml
3. http://www.dmoz.org

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

where i is the position of the retrieved document d in the list, and rel_i = 1 if the ith document belongs to facet v and 0 otherwise.

This score gives a measure of the contribution of a particular facet to the top-ranked results for a particular query. Consider a case of two facets: The first facet has 5 documents all of which are in the top ten results for a query. The second has 5 documents but whose ranks are above 30. Although the two facets have the same number of documents, the first facet has more important results for the query and this is reflected in the $importance$ score.

2.2. *Computing distances between nodes*

Distances between nodes could be measured by various metrics like the number of edges between the nodes and the inverse of the cosine similarity between the topics corresponding to the nodes. Cabanac et al. [CAB 10] also propose a distance metric between documents organized in a hierarchy. In this paper, we propose a weighted distance scheme that reflects the semantic distances between topics.

Since the basic relations in the taxonomy are the parent-child relations, distance between any two nodes are represented using the connection weights between the parent-child pairs associated. In taxonomy **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:

$$D(v_i, v_j) = 2^{-l} \quad [2]$$

The connection weights between the nodes are assumed to be lower in the base of the hierarchy and become progressively larger as one moves up. The reasoning behind this weighing scheme is to highlight the fact that semantic differences between nodes are more prominent in the top of the hierarchy. Using this metric, the distance between two nodes v_m and v_n in a taxonomy tree **T** is defined as the sum of the connection weights between all the nodes v_x that span the path between v_m and v_n. Figure 1 illustrates the weighing scheme used.

3. Selecting the set of presentation facets

The task of selecting the set of presentation nodes is defined as follows: Given the taxonomy tree **T** of e.g., the DMOZ hierarchy and the set of *activated* nodes **V** in **T**, a subset of k nodes in **T** that are representative of the search results **R** relevant for query q is computed. k is chosen based on the size of the interface medium and the cognitive load acceptable for a user. The idea is that when the number of relevant facets to show is large, there is a need for a kind of summarization over the facet nodes. In contrast, when the number of facets to show is small, a user might be interested to see more related facets, so we would like to expand the facets.

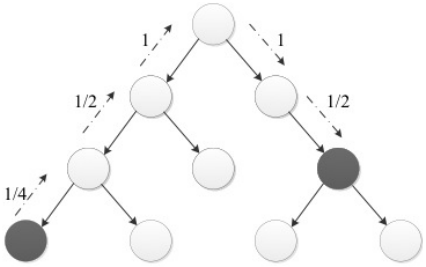

Figure 1. *Distances between two activated (shaded) nodes is* $1/4+1/2+1+1+1/2$. *Weights adjacent to dotted lines denote distances*

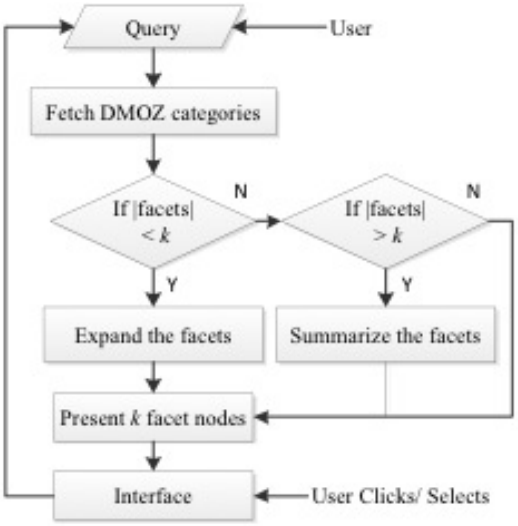

Figure 2. *Schematic overview of the selection of facets*

Figure 2 shows a flowchart of the facet selection, which includes the two basic building blocks - summarization of the facets (section 3.1) and expansion of the facets (section 3.2). When the user presents a query, the DMOZ facets associated with the query result pages are first extracted, i.e., the *activated nodes* are identified. Next, if the number of *activated* facets associated with the query is greater 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 algorithms are discussed below.

3.1. *Summarizing the facets*

3.1.1. *Model S1: Subtree density*

This model finds nodes which represent dense clusters of facets, where each facet has many search results important for the query. The subtree $\mathbf{S_V}$ associated with a node v is the set containing the node and its descendants in $\mathbf{T}$ as defined in Eq. 3:

$$\mathbf{S_V} = \{s|s \in descendants(v)\} \quad [3]$$

where descendants(v) is a recursive function that computes the children, children of the children etc. until the last level of node v.

Subtrees are built in this manner for all the *activated* nodes identified for a search query. However, for some queries, DMOZ could activate not only the lowest level facets, but also some of their ancestors. For example, assume the facets Kids_and_-Teens/Entertainment/Animation and Kids_and_Teens/Entertainment are both *activated*. Note that one of them is a descendant of the other. In such a case, while expanding subtrees of the two nodes, there could be many overlapping nodes. That is, some nodes could be part of more than one subtree. To present as many different facets as possible, it is desirable to have subtrees that are disjoint. Therefore, out of every pair of nodes (for which we build subtrees) if one is a descendant of the other, we choose only one of the nodes. In this case, the descendant lowest in the tree is chosen. This ensures that we do not have overlapping subtrees, given that the nodes are all mapped to a tree (hierarchy).

Then, the 'density' of each subtree (**S**) is computed using Eq 4:

$$density(\mathbf{S}) = \frac{\sum_{v \in S} importance(v, \mathbf{R})}{|\mathbf{S}|} \quad [4]$$

where $importance(v, \mathbf{R})$ is computed as in Eq. 1 and $|\mathbf{S}|$ is the size of the subtree in terms of number of nodes $v \in \mathbf{S}$.

Due to the limitation on the number of nodes that can be presented in the interface and since there could be a large number of subtrees, it is essential to choose a representative node for each subtree. One possible representative is the medoid identified as the node that has the minimum average distance to all the other nodes of the subtree. The distances between nodes in the subtree are computed using the distance metric (Sec. 2.2). Every medoid, m, is then assigned the score computed using Eq. 5:

$$score(m) = proximity(m, \mathbf{S}) * density(\mathbf{S}) \quad [5]$$

$proximity(m, \mathbf{S})$ is the inverse of the distance between the medoid and the root of the subtree **S**, and $density(\mathbf{S})$ is the subtree density computed in Eq. 4. 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.

The medoids are ranked based on Eq. 5, and the best k medoids, i.e. the ones of the clusters with the highest score, are then presented. Model S1 thus promotes relevance of a facet via the importance score that reflects the ranking of the documents belonging to that facet, and aims at representing dense clusters of search results.

3.1.2. *Model S2: Selection by facet importance and share*

Agrawal et al. [AGR 09] have proposed an algorithm that re-orders the search results (documents) based on two attributes - probability of the query belonging to a facet and the quality value of the search result (document) for the query, interpreted

as the likelihood of the document satisfying the user intent given the query. Inspired by this approach, we propose a metric that ranks facets based on the importance score and the volume or share of search results of the facets. The share of a facet for a query is computed as the ratio of results belonging to the facet.

$$value(v, q) = share(v, q) * importance(v, \mathbf{R}) \quad [6]$$

where v is a facet, q is the query and importance is estimated using Eq. 1. Model S2 promotes the importance of a facet based on the amount and ranking of documents that belong to it.

3.1.3. *Model S3 k-medoid clustering*

Another approach to choose the facets of the query's results is by k-medoid clustering of the *activated* nodes. This clustering is performed using a matrix of pairwise distances between nodes, computed using the distance metric in Sec. 2.2. The k medoids obtained in this way are suitable representatives [OSI 05] of the (numerous) facets that are *activated* by query q and form the set of nodes presented to the user. Model S3 promotes diversity of the results.

3.1.4. *Model S4: Parent facets subsuming all the activated facets*

Another solution to the problem of choosing the right facets could be to display a chosen number (k) of facets that are parents or ancestors to the *activated* nodes, and that are at the same level in the hierarchy. So, these are the nodes that are most specific, but still encompass all facets to which the relevant documents belong. Although some information could be lost, as with any kind of summarization, it enables different broad aspects to be presented when the interface cannot allow all the *activated* nodes. Model S4 basically promotes generalization of the search results.

3.2. *Expanding the facets*

3.2.1. *Model E1: Expansion with parent nodes*

The model uses the parents of the *activated* nodes as possible facets. The set of nodes **Q'** used for recommending facets is defined in Eq. 7:

$$\mathbf{Q'} = \{s | s = P(v)\ AND\ s \notin \mathbf{V}\} \quad [7]$$

where **V** denotes the set of *activated* facet nodes associated with the query, v is any node in **V** and $P(v)$ is a function that fetches the parent of the node v in the hierarchy. Expanding an *activated* node with its direct parent is related to model S4 as model E1 basically promotes a limited generalization of the search results.

3.2.2. *Model E2: Expansion with sibling nodes*

An intuitive way of expanding the *activated* nodes is by considering the sibling nodes indicated in Eq. 8:

$$\mathbf{Q'} = \{s | s \in B(v)\ AND\ s \notin \mathbf{V}\} \quad [8]$$

Query	Facets	Model
Arts	'Painting', 'Music', 'Photography', 'Renaissance', 'History', 'Dance', 'Organizations'	S1
Animal	'Pets', 'Mammals', 'Aquaria', 'Conservation', 'Biodiversity', 'Endangered Species', 'Organizations'	S2
Atom	'Atomic Kitten', 'Atoms, Nuclei & Particles', 'Chemistry', 'Elements', 'Periodic_Tables'	BL
Electricity	'Electricity & Magnetism', 'Energy', 'Renewable Energy', 'Safety', 'Lightning', 'Electric Power'	S3
Medicine	'Conditions & Diseases', 'Careers', 'First Aid', 'Safety', 'Science', 'Emotional Health & Wellbeing', 'Substance Abuse'	S4
Cricket	'Sports and Hobbies', 'Animals', 'Insects'	E1
Cricket	'Baseball', 'Volleyball', 'Football', 'Swimming & Diving', 'Dragonflies', 'Beetles', 'Butterflies & Moths'	E2
Hubble's Telescope	'Astronomy & Space', 'Solar System', 'Edwin Hubble', 'Galleries', 'Sun', 'Earth', 'Asteroids, Comets, & Meteors', 'Black Holes', 'Astronomers'	E3

Table 1. *Examples of facets chosen*

where $B(v)$ is a function that fetches the siblings of a node v. Model E2 promotes related facets that are not directly related to the query.

3.2.3. *Model E3: Expansion with parent nodes and filtering by subtree density*

In this model, we first select the parent node of an *activated* node, but then refine the presentation by applying the subtree density model (Sec. 3.1.1) on the parent node. This application yields a medoid of the subtree by which the facets could be expanded. Model E3 finds related facets and selects them from clusters with high density of relevant documents.

Table 1 shows examples of facets chosen by the different models.

4. Experiments and Results

4.1. *Experimental setup*

Two sets of queries have been used for evaluation. The first set of queries (**Q1**) contains titles of English Wikipedia articles. These are relatively clean and well-formatted queries. The second set of queries (**Q2**) comprises real user queries collected by Torres et al. [DUA 10]. These are queries for children extracted from the AOL query log.

In either case, 1200 queries were selected and submitted to the Bing search engine, restricting the search results to the Web pages from the DMOZ Kids and Teens

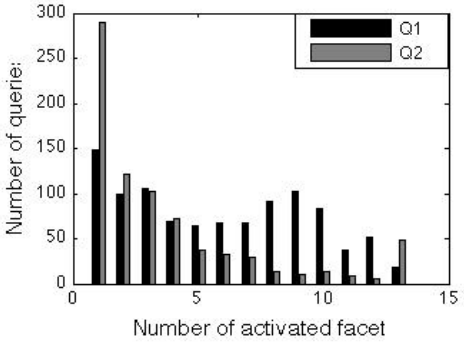

Figure 3. *Number of activated nodes for the queries in* **Q1** *and* **Q2**

	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

Table 2. *Statistics of the query sets used in evaluation of the summarization and expansion*

subdirectory. The search results are also annotated with subject categories. The facets chosen by all the models for each query of the two query sets were presented to five Crowdflower [4] evaluators, who were asked to judge the relevance and diversity of these facets. Only queries for which the agreement among Crowdflower evaluators was over 80% (as reported by Crowdflower) were retained. Table 2 shows the number of queries used for evaluation of the models. Figure 3 shows the distribution of *activated* facet nodes over the query results. The presentation nodes will be selected based on these sets.

4.2. *Evaluating the summarization and expansion of the facets*

We evaluate two important aspects: relevance of the selected facet for the query and diversity of the aggregated k facets in the presentation set. The baseline (BL) model for summarization uses the top k distinct *activated* nodes from the ranked search results that were returned by the search engine.

Relevance of the selected facets was evaluated by comparing with the ground truth judgements for each method based on the precision. The precision of the five models for summarization and of the three models for expansion is shown in figure 4. The number of facets presented (k) is shown on the X-axis. Recall is not computed because the evaluation of the full taxonomy of more than 7000 facets by Crowdflower evaluators for each of the queries is not realistic and because of the restricted number of facets that can be shown to the user. Overall the subtree density model (S1) performs best when considering the relevance of the facets for the query. This makes sense as

4. http://crowdflower.com/

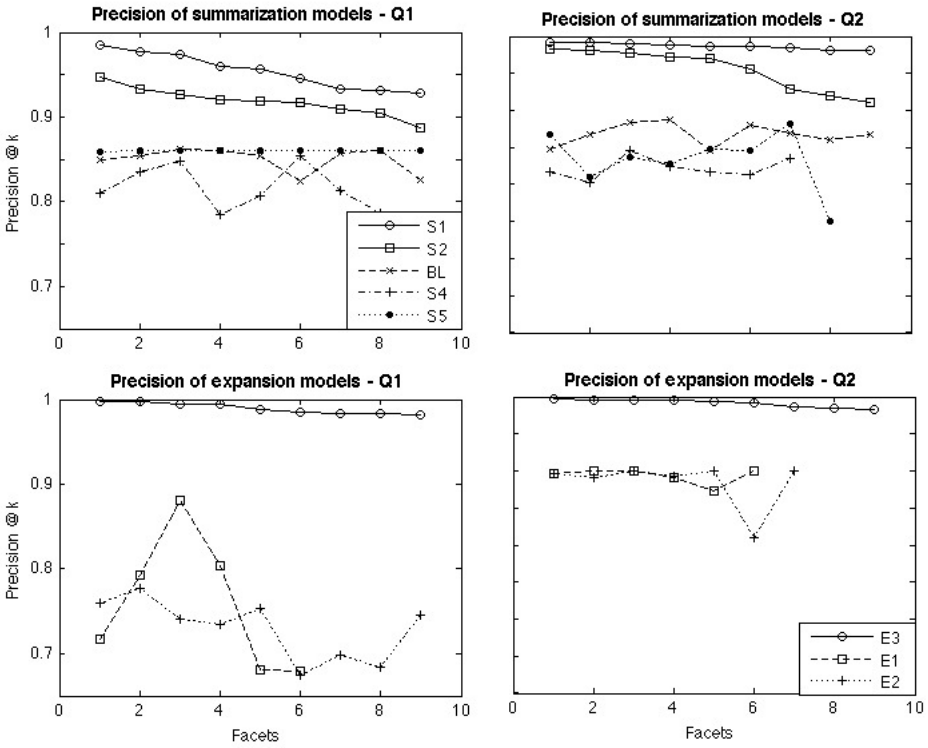

Figure 4. *Precision @ k of the summarization and expansion models for the first nine facets for query sets* **Q1** *and* **Q2**

this metric most strongly emphasizes that a representative facet should come from a dense cluster of facets in the taxonomy, where on average the facets of the cluster represent many important search results. Model S2 behaves similarly but exhibits a decreased performance. This can be explained by its lack of the density component that promotes facets in dense regions. The k-means method (S3) behaves in an unstable way with regard to relevance. With this method focussing on diversity, outliner clusters might be selected which might contain irrelevant results. Finally, selecting parent facets that subsume the search results (S4) forms a weak baseline as it misses density information of the facets and importance information of the search results. Also, it might promote irrelevant facets through the generalization process.

The results of facet expansion are also indicated in figure 4. The results of expanding with parents (E2) and siblings (E3) are comparatively weak. This is because we lack additional selection criteria. These are incorporated in the expansion model that expands with parents and their descendants (E1), but then selects a representative facet based on density of the facets that contain search results and importance of the search

	S1 Q1	S1 Q2	S2 Q1	S2 Q2	BL Q1	BL Q2	S3 Q1	S3 Q2	S4 Q1	S4 Q2
Rank1	79.86	94.84	81.90	98.76	63.80	90.93	81.90	95.26	59.96	86.80
Rank2	4.07	2.89	2.04	0.62	2.04	1.65	4.07	2.47	2.26	3.30
Rank3	11.99	1.65	11.99	0.41	23.98	4.12	9.95	0.62	5.88	1.65
Rank4	2.04	0.41	2.04	0.21	7.92	2.06	2.04	0.41	2.04	2.89
Rank5	2.04	0.21	2.04	0	2.26	1.24	2.04	1.24	29.86	5.36
Total	100	100	100	100	100	100	100	100	100	100

Table 3. *Diversity of facets produced by the summarization models (percentage) for query sets* **Q1** *and* **Q2**

results. We select nodes that are situated in dense regions of the facets *activated* by the query that represent important search results.

Another important evaluation, now of the whole presentation set, regards the diversity of the set of selected facets when representing the query results (in the case when we summarize the set of facets). To evaluate the diversity of the facets, we put together five clusters of related facets (that were judged relevant by Crowdflower evaluators)- one for each summarization model, per query for the 508 queries from **Q1** and 523 queries from **Q2**. We asked five Crowdflower evalutors 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 most diverse cluster. Only queries where annotators showed an agreement of more than 80% were used, leaving us with 441 queries from **Q1** and 485 queries from **Q2**. The results on these queries (in percentages) are given in Table 3.

Firstly note that for all the five models, majority of the queries' facets are ranked at 1, indicating that most models already show facets that are very diverse. Second, as far as diversity is concerned, S2 and S3 have comparable performance. As expected, the k-medoid clustering does promote diversity. S1 performs quite well in terms of diversity, but this is not the case for S4. The reason for the relatively lower scores of S4 could be that the categories come closer to each other as one moves up the hierarchy, which also justifies our weighted distance scheme.

5. Future work and Conclusions

In this paper we have proposed several approaches for summarizing and expanding search results mapped to a subject taxonomy. The methods have been evaluated based on relevance (measured by precision @ k) and diversity. The subtree density model is the most successful method for both summarization and expansion as it optimizes both relevance and diversity.

A next step in our research is to develop navigation models for interactive browsing using the presented facets and their corresponding Web pages. Also, while these

models have been designed for a topical taxonomy, we would like to evaluate them on other kinds of taxonomies like a stylistic, temporal and opinion oriented taxonomy.

6. References

[AGR 09] AGRAWAL R. GOLLAPUDI S. H. A., S L., "Diversifying search results", *Proceedings 2nd ACM International Conference on Web Search and Data Mining*, Barcelona, Spain, 2009, ACM, p. 5-14.

[AHM 10] AHMED-OUAMER R., HAMMACHE A., "Ontology-based information retrieval for e-Learning of computer science", *Machine and Web Intelligence (ICMWI), 2010 International Conference on*, oct. 2010, p. 250 -257.

[CAB 10] CABANAC G., CHEVALIER M., CHRISMENT C., JULIEN C., "Organization of digital resources as an original facet for exploring the quiescent information capital of a community", *International Journal on Digital Libraries*, Springer, 2010, p. 1–23.

[CHE 00] CHEN H., DUMAIS S., "Bringing order to the web: Automatically categorizing search results", *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 2000, p. 145–152.

[DOU 11] DOU Z., HU S., LUO Y., SONG R., WEN J., "Finding dimensions for queries", *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, ACM, 2011, p. 1311–1320.

[DUA 10] DUARTE TORRES S., HIEMSTRA D., SERDYUKOV P., "An analysis of queries intended to search information for children", *Proceedings of the Third Symposium on Information Interaction in Context*, ACM, 2010, p. 235–244.

[JÄR 00] JÄRVELIN K., KEKÄLÄINEN J., "IR evaluation methods for retrieving highly relevant documents", *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, 2000, p. 41–48.

[LI 10] LI C., YAN N., ROY S., LISHAM L., DAS G., "Facetedpedia: dynamic generation of query-dependent faceted interfaces for Wikipedia", *Proceedings of the 19th International Conference on World Wide Web*, ACM, 2010, p. 651–660.

[OSI 05] OSINSKI S., WEISS D., "A concept-driven algorithm for clustering search results", *Intelligent Systems, IEEE*, vol. 20, num. 3, 2005, p. 48–54, IEEE.

[SUS 08] SUSHMITA S., LALMAS M., TOMBROS A., "Using digest pages to increase user result space: Preliminary designs", *SIGIR 2008 Workshop on Aggregated Search*, 2008.

[YEE 03] YEE K., SWEARINGEN K., LI K., HEARST M., "Faceted metadata for image search and browsing", *Proceedings of the SIGCHI Conference on Human factors in Computing Systems*, ACM, 2003, p. 401–408.

[ZAM 99] ZAMIR O., ETZIONI O., "Grouper: a dynamic clustering interface to Web search results", *Computer Networks*, vol. 31, num. 11, 1999, p. 1361–1374, Elsevier.