Automatic Discovery of Useful Facet Terms

Wisam Dakka
Columbia University
wisam@cs.columbia.edu

Rishabh Dayal
Columbia University
rd2214@columbia.edu

Panagiotis G. Ipeirotis
New York University
panos@nyu.edu

ABSTRACT

Databases of text and text-annotated data constitute a significant fraction of the information available in electronic form. Searching and browsing are the typical ways that users locate items of interest in such databases. Faceted interfaces represent a new powerful paradigm which has been proven to be a successful complement to keyword searching. Thus far, the generation of faceted interfaces relied either on manual identification of the facets, or on *apriori* knowledge of the facets that can potentially appear in the underlying database. In this paper, we present our ongoing research towards automatic identification of facets that can be used to browse a collection of free-text documents. We present some preliminary results on building facets on top of a news archive. The results are promising and suggest directions for future research.

1. INTRODUCTION

A significant amount of information is available in electronic form and stored in online databases. Users who want to locate information in online databases typically rely on one of the two major paradigms: they either use a direct, keyword-based search, or they browse through the contents of the database to locate items of interest. Commonly, browsing is supported by a single hierarchy or a taxonomy that organizes thematically the contents of the database. Unfortunately, a single hierarchy can very rarely organize coherently the contents of a database. For example, consider an image database. Some users might want to browse by style, while other users might want to browse by topic.

An alternative to single, monolithic hierarchies is to use multiple, *faceted hierarchies* [10] for browsing. Pollitt [24] and Yee et al. [29] showed that faceted hierarchies are superior than single, monolithic hierarchies. In faceted interfaces, users can zoom, using one dimension, to a certain point in the hierarchy, and then slice the database and switch browsing to another hierarchy. For example, consider the case of a user looking for images of children playing with dogs in a farm. Having multiple hierarchies, the user can browse first through the hierarchy "Animals" and select the category "Mammals → Carnivores → Dogs." Then, having the "Animal" dimension fixed, the user can browse through the hierarchy "Places" to locate images with farms, and then browse through the hierarchy "People" to locate images with children. Such multifaceted interfaces expose the contents of the underlying database and can help users more quickly locate items of interest.

So far, the systems that use faceted interfaces are built manually. One of the fundamental tasks required to allow wide deployment of faceted interfaces is to build techniques for *automatic construction of faceted interfaces*. Building a multifaceted interface on top of a database consists of two main steps:

- Identifying the facets that are useful for browsing the underlying database, and
- Building a hierarchy for each of the identified facets.

In our previous work [4], as part of our effort to fully automate the construction of faceted interfaces, we introduced a *supervised* approach for extracting useful facets from a database of text or text-annotated data. Our technique (briefly described in Section 2) relies on WordNet hypernyms and on a Support Vector Machine (SVM) classifier to assign new keywords to facets. After training, our algorithm automatically assigns new keywords to the appropriate facets and then discovers the important facets that appear in a database of text-annotated objects.

Unfortunately, the algorithm had some limitations. First, since we relied on a supervised learning technique, the facets that could be identified by our algorithm were, by definition, limited to the facets that appeared in the training set. Second, since the algorithm relied on WordNet [5] *hypernyms*[1], it was difficult to apply our technique on objects annotated with named entities (or even noun phrases), since WordNet has rather poor coverage of named entities. Finally, while our algorithm generated high quality faceted hierarchies from databases of keyword-annotated objects, the quality of the respective hierarchies built on top of text documents (e.g., news articles) was comparatively low.

In this paper, we present some preliminary results of our ongoing research that aims to alleviate the shortcomings of our previous work. Our goal is to create techniques that fully automate the extraction of the useful facets from free-text. In particular, our goals are to:

1. automatically discover, in an *unsupervised* manner, a set of candidate facet terms from free text;
2. automatically group together facet terms that belong to the same facet;
3. build the appropriate browsing structure for each facet.

[1]Hypernym is a word whose meaning includes the meanings of other words, as the meaning of vehicle includes the meaning of car, truck, motorcycle, and so on.

SIGIR '06 Seattle, Washington USA

The basic intuition behind our approach is that high-level facet terms rarely appear in the database documents. For example, consider the named entity "*Jacques Chirac*." This term would appear under the facet "*People → Political Leaders*." Furthermore, this named entity also implies that the document can be potentially classified under the facet "*Regional → Europe→ France*." Unfortunately, these (facet) terms are not guaranteed to appear in the original text document. However, if we "expand" the named entity "*Jacques Chirac*" using an external resource, such as Wikipedia, we can expect to encounter these terms more frequently. Our hypothesis is that facet terms will emerge after the expansion and their frequency rank will increase in the new, expanded database.

The rest of the paper is structured as follows. In Section 2, we give the necessary background. Then, in Section 3, we discuss in detail our ongoing work for unsupervised identification of facets and facet terms, and, in Section 4, we report some initial experimental results. Finally, in Section 5, we review related work and, in Section 6, we discuss future work and conclude the paper.

2. BACKGROUND

While work on automatic construction of *faceted* interfaces is relatively new, automatic creation of subject hierarchies has been attracting interest for a long time, mainly in the form of *clustering* [3, 19, 30]. However, automatic clustering techniques generate clusters that are typically labeled using a set of keywords, resulting in category titles such as "*battery california technology mile state recharge impact official hour cost government*" [11]. While it is possible to understand the content of the documents in the cluster from the keywords, this presentation is hardly ideal.

An alternative to clustering is to generate hierarchies of *terms* for browsing the database. Sanderson and Croft [27] introduced the subsumption hierarchies and Lawrie and Croft [14] showed experimentally that subsumption hierarchies outperform lexical hierarchies [21, 22, 23]. Kominek and Kazman [12] use the hierarchical structure of WordNet [5] to offer a hierarchy view over the topics covered in videoconference discussions. Stoica and Hearst [28] also use WordNet together with a tree-minimization algorithm to create an appropriate concept hierarchy for a database.

All these techniques generate a *single* hierarchy for browsing the database. In [4], we presented a supervised technique for separating the terms into different facets, before building a hierarchy for each facet. For example, we put the words "*cat*" and "*dog*" are under the "*Animals*" facet, while we put the words "*mountain*" and "*fields*" under the "*Topographic Features*" facet. To segment the terms into facets, we used existing databases that have metadata organized across different facets (e.g., the Corbis image database, in which each of the annotation keywords has an associated facet).

Using such data, we use the facet as a target class and the keywords as features, in order to assign keywords to the appropriate facet. To allow our technique to generalize, we rely on the observation that keywords under the same facet tend to have similar hypernyms. Based on this observation, we expand each keyword using its hypernyms from a lexical corpus, such as WordNet [5]. After the expansion, each keyword is represented as a set of words. For example, the word "*cat*" is represented as "*cat, feline, carnivore, mammal, animal, living being, object, entity*". The new representation allows the classifier to generalize more easily and assign unseen words to the correct facets.

One of the disadvantages of this algorithm is its supervised nature. While we expect the algorithm to perform well on databases that have facets similar to the ones in the Corbis dataset, the algorithm, by definition, cannot discover facets that did not appear in the training set. Also, the algorithm cannot work well with terms that do not appear in WordNet, thus hinting that another form of expansion might be necessary. Next, we describe our approach for overcoming these problems.

Facets
Location
Institutes
History
People
Social Phenomenon
Markets
Nature
Event

Figure 1: Facets identified by human annotators in a small collection of 100 news articles from The New York Times.

3. AUTOMATIC FACET DISCOVERY

In this section, we describe our research-in-progress for automatic discovery of useful facet terms from free-text documents. Section 3.1, motivates our work and gives an overview of our approach. Then, Section 3.2 describes in detail our approach, and Section 3.3 describes our current work in progress.

3.1 Overview and Motivation

We are interested in providing a multifaceted interface for the news archive of the Newsblaster[2] project [18]. The Newsblaster archive contains news articles from 24 English news sources, dating back to 2001. Searching and accessing a big news archive is often a hurdle for news reporters and researchers. As part of our efforts to allow easier access to the Newsblaster archive, we are working towards creating a multifaceted interface on top of the archive, which will automatically adapt to the contents of the underlying news collection. A significant component of this system is the ability to automatically discover the facets that can be used to browse the archive.

Initially, to identify the facets in the collection, we ran a small pilot study. We picked randomly a hundred stories from *The New York Times* archive in Newsblaster and we asked annotators to manually assign each story to several facets that they considered appropriate and useful for browsing. The most common facets identified by the annotators were "*Location*," "*Institutes*," "*History*," "*People*," "*Social Phenomenon*," "*Markets*," "*Nature*," and "*Event*." For these facets, the annotators also identified other "sub-facets" such as "*Leaders*" under "*People*" and "*Corporations*," under "*Markets*."

From the results of the pilot, we also noticed one clear phenomenon: the terms for the useful facets do not usually appear in the news stories. Typically, journalists do not use general terms, as those used to describe facets, in their stories. For example, a journalist writing a story about *Jacques Chirac* will not necessarily use the term "*Political Leader*" or the terms "*Europe*," or even "*France*." Such (missing) terms are tremendously useful for identifying the appropriate facets for the story.

After conducting this informal experiment, it became clear that a tool for automatic discovery of useful facet terms should exploit some external resource that could return the appropriate facet terms. Such an external resource should provide the appropriate context for each of the terms that we extract from the database. As external resources for providing context we used WordNet, Google, and Wikipedia. The basic idea is to query these resources and ex-

[2] http://newsblaster.cs.columbia.edu

amine which terms tend to co-occur often with the terms from the database. Our algorithm assumes that facet terms are rare terms in the original database but co-occur frequently in the external resources with the terms that appear in the original database. As a result, a key step of our approach is an expansion procedure, in which the important terms from each news story are expanded with "context terms" derived from the external resources. The expanded documents then contain many of the terms that can be used as facets. Next, we describe our algorithm in detail.

3.2 Algorithm

To identify the candidate facet terms, we identify terms that were rather infrequent in the original database, but are frequent in the database with the expanded documents. In particular, our algorithm proceeds as follows:

1. For each document in the database, identify the important terms that are useful to characterize the contents of the document.

2. For each term in the original database, query the external resource and retrieve the terms that appear in the results. Add the retrieved terms in the original document, in order to create an expanded, "context-aware" document.

3. Analyze the frequency of the terms, both in both the original and the expanded database and identify the candidate facet terms.

Next, we describe in details each of the steps of the algorithm.

Step 1: Identifying important terms in a document: Typically, the named entities mentioned in a news story are terms that give important clues about the topic of the document. This is further reinforced by existing research (e.g., [9, 6]) that shows that the use of named entities increases the quality of clustering and improves news event detection. We built on these ideas and use the named entities extracted from each news story as descriptions of the important aspects of the document. Furthermore, we use the "Yahoo Term Extraction"[3] web service, which takes as input a text document and returns a list of significant words or phrases extracted from the document. We use this service as a second tool for identifying important terms in the document.

Step 2: Deriving Context from External Resources: Let $O = \{o_1 \dots o_n\}$ be the original database of news stories and let A be the available operators that identify important terms in the original news stories.[4] We define as $A(o_i)$ the set of terms that are extracted from the document o_i after using all the operators in A. Finally, we denote with $W = \{w_1 \dots w_m\}$ the available external resources. (Currently, we use Wikipedia, Google, and WordNet.) We define as $E(A(o_i), w_k)$ the expanded, "context-aware" set of terms that characterize the document o_i, after expanding each term in $A(o_i)$ using the external resource w_k. To generate the set $E(A(o_i), w_i)$, we query the resource w_k with each term in $A(o_i)$, and we extract the terms that appear in the results returned by w_k.

For example, consider a document o that discusses the actions of *Jacques Chirac* during the *2005 G8 summit*. In this case, the set $A(o)$ may contain the terms

$$A(o) = \{\textit{Jacques Chirac}, \textit{2005 G8 summit}\}$$

Assume that we use Wikipedia as the external resource; we query Wikipedia with the two terms in $A(o)$ and we analyze the returned results. From the documents returned by Wikipedia, we identify additional context terms for the two terms in the original $A(o)$: the term *president of France* for the original term *Jacques Chirac* and the terms *Africa debt cancellation* and *global warming* for the original term *2005 G8 summit*. Therefore, the set $E(A(o), \textit{Wikipedia})$ contains the two original terms in $A(o)$ and the three additional context terms, *president of France*, *Africa debt cancellation*, and *global warming*.

At the end of this process, for each external resource w_j we expand the original documents O to create a new collection E. Each document c_i in this collection is the concatenation of the original document $A(o_i)$ and $E(A(o_i), w_i)$.

Step 3: Term Frequency Analysis: In this step, we identify terms that are good candidates for facet terms. Our algorithm is based on the intuition that facet terms are infrequent in the original database, but frequent in the expanded one. To measure the difference in frequency, we define the next two functions:

- **Frequency-based Shifting**: For each term t, we compute the frequency difference as

  $$\textit{Shift}_f(t) = \textit{Freq}_E(t) - \textit{Freq}_O(t)$$

  where $\textit{Freq}_E(t)$ and $\textit{Freq}_O(t)$ are the frequencies of t in the expanded and the original database, respectively. Due to the Zipfian nature of the term frequency distribution, this function tends to favor terms that have already high frequencies in the original database. Terms with high frequencies demonstrate higher increases in frequency, even if they are less popular in the expanded database compared to the original one. The inverse problem appears if we use ratios instead of differences. To avoid the shortcomings of this approach, we introduce a rank-based metric that measures the differences in the ranking of the terms.

- **Rank-based Shifting**: We assume that we have a "bucket" function B that assigns terms to bins based on their ranking. In this paper, we use the function

  $$B(t) = \lceil \log_2(\textit{Rank}(t)) \rceil$$

  where $\textit{Rank}(t)$ is the rank of the term t in the database. After computing the bin $B_O(t)$ and $B_E(t)$ of each term t in the original and expanded database, respectively, we define the shifting function to be

  $$\textit{Shift}_r(t) = B_O(t) - B_E(t)$$

In our approach, a term becomes a candidate facet term only if both $\textit{Shift}_f(t)$ and $\textit{Shift}_r(t)$ are positive. We report some initial results in Section 4.

3.3 Work in progress

We are currently working on grouping together the candidate facet terms. Our approach is to build subsumption hierarchies [27] on top of the extracted candidate facet terms, and keep the high-level categories of the hierarchy as independent facets. We plan to contrast the results with the hierarchies created using the non-expanded database. Furthermore, we will present comparative results across the facets generated using various expansion resources. For example, the generated facets are expected to be different when we use WordNet for the expansion (in a spirit similar to Stoica and Hearst [28]) compared to the case where we use Wikipedia as the

[3] `http://developer.yahoo.com/search/content/V1/termExtraction.html`

[4] We currently use the LingPipe named entity tagger and the Yahoo! Term Extraction web service.

t	$Shift_f(t)$	$Shift_r(t)$
state	1960	1
president	1799	1
nation	760	1
urban	716	2
executive	713	1
metropolitan	707	3
geographical	528	2
capital	459	1
organization	457	1
area	335	1
world	288	1
region	283	1
disease	277	1
person	269	1

Figure 2: Some indicative facet terms in the database expanded using WordNet, as identified by our algorithm.

year, new, time, people, state, work
school, home, mr, report, game, million
week, percent, help, right, plan, house
high, world, american, month, live, call, thing

Figure 3: Facet terms identified by a simple subsumption-based algorithm [27], without using the expansion algorithm.

external resource. Another alternative that we would like to investigate is to cluster all the candidate facet terms, and name each cluster. The names will be the our facets.

Next, we present some preliminary experimental results showing that our algorithm indeed identifies terms that are good for generated faceted hierarchies.

4. EXPERIMENTAL RESULTS

Now, we report some initial experimental evaluation for our algorithm of Section 3. We first describe briefly our data set and then present some preliminary results.

Data Set: Our data set contains 1,700 news stories from one day of November 2005, as retrieved from 24 news sources. We processed each document in the collection using the LingPipe named entity tagger and we used the Yahoo! Term Extractor to identify additional important terms in each news story. We then expanded the terms using WordNet hypernyms as the external resource to create the expanded database. Finally, we computed the metrics $Shift_f(t)$ and $Shift_r(t)$ for each of the terms in the databases.

Results: In Figure 2, we present the top candidate terms identified by our algorithm from the WordNet-expanded collection. These are the terms with the highest values of $Shift_f(t)$ and $Shift_r(t)$. We believe that the majority of these terms are good candidates to generate facet hierarchies. The terms *state*, *urban*, *metropolitan*, *geographical*, *capital*, *area*, *world*, and *region* are clearly terms that could populate a facet that allows navigation through the "*Location*" facet, identified by the annotators during the pilot study (in a subset of the data set). Similarly, the terms *president*, *executive*, and *person* could populate the "*People*" facet, also identified by the annotators during the pilot study. We are currently working on techniques that would group together these terms under the appropriate facet.

As a comparison, we used the subsumption algorithm from [27] on the original documents, to identify the top-level hierarchy terms for the given data set. In Figure 3, we list the best terms as identified by the subsumption algorithm. While there are some isolated terms that can be used for facet generation, many of the terms are not useful for faceted browsing.

5. RELATED WORK

Faceted interfaces, which use multiple, orthogonal classification schemes to present the contents of a database, become increasingly popular. A large number of e-commerce web sites use faceted interfaces [13], based on engines provided by companies such as Endeca[5] and Mercado,[6] which expose the facets that are already defined for the products (e.g., "by price," "by genre" and so on). Systems developed in academia, such as HiBrowse [24], OVDL [17], and Flamenco [29], demonstrate the superiority of faceted interfaces over single hierarchies. Our work on automatic construction of multifaceted interfaces contributes to this area and facilitates the deployment of faceted databases. In an orthogonal direction, Ross and Janevski [25] present work on *searching* faceted databases and describe an associated entity algebra and a query engine.

As an alternative to creating a separate hierarchy for each collection, Chaffee and Gauch [1] presented a system that uses a personalized ontology to offer a common browsing experience across collections of web pages (i.e., web sites) that organize their contents in different ways. Other, less common browsing structures were proposed (e.g., wavelet-based text visualization [20], dynamic document linking [8]) but hierarchy-based approaches continue to be the most popular interfaces for faceted browsing.

Our approach on identifying facet terms is conceptually similar to the *skew divergence* of Lee [15], which is used to identify substitute terms (e.g., that "fruit" can approximate "apple" but not vice-versa). Recent work by Sahami and Heilman [26] tries to identify semantically similar text snippets (e.g., "UN Secretary-General" and "Kofi Anan"), and could also be useful in our scenario where we are trying to identify generic facet terms that subsume the important terms that appear in our documents. On a broader context, our work relies on *distributional analysis* [15] of two collections (the original and the expanded one) to identify terms that have high distributional differences across the two collections, hoping that these terms are good facets terms. Distributional analysis has also been used by Gabrilovich et al. [7] for novelty detection in a stream of news, and by Cronen-Townsend et al. [2] to measure the "clarity" of a query with respect to a given document collection.

6. CONCLUSIONS AND FUTURE WORK

We presented a method for automatically identifying terms that are useful for building faceted hierarchies. Our techniques build on the idea that external resources, when queried with the appropriate terms, provide useful context that is valuable for locating the facets that appear in a database of text documents. Our initial experimental results over a small news archive are encouraging and show directions for future research.

One of our immediate plans is to create techniques that can group together terms that are useful for facet browsing: this will result in a coherent set of facets for the underlying database. We also plan to examine different techniques for extracting useful terms from the documents, such as topic signatures [16]. Furthermore, we plan to examine the quality of the generated facet terms when using Wikipedia and Google as external resources. Finally, we plan to evaluate the automatically generated facets in a task-oriented eval-

[5] http://www.endeca.com
[6] http://www.mercado.com

uation with human users, to examine what are the main shortcomings of automatically generated facets when compared with manually created faceted interfaces.

7. REFERENCES

[1] J. Chaffee and S. Gauch. Personal ontologies for web navigation. In *Proceedings of the 2000 ACM Conference on Information and Knowledge Management (CIKM 2000)*, pages 227–234, 2000.

[2] S. Cronen-Townsend, Y. Zhou, and W. B. Croft. Predicting query performance. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2002*, pages 299–306, 2006.

[3] D. R. Cutting, D. R. Karger, J. O. Pedersen, and J. W. Tukey. Scatter/Gather: A cluster-based approach to browsing large document collections. In *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'92*, pages 318–329, 1992.

[4] W. Dakka, P. G. Ipeirotis, and K. R. Wood. Automatic construction of multifaceted browsing interfaces. In *Proceedings of the 2005 ACM Conference on Information and Knowledge Management (CIKM 2005)*, pages 768–775, 2005.

[5] C. Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press, May 1998.

[6] E. Filatova and V. Hatzivassiloglou. Marking atomic events in sets of related texts. In *Recent Advances in Natural Language Processing, Part III*, pages 247–256, 2003.

[7] E. Gabrilovich, S. Dumais, and E. Horvitz. Newsjunkie: Providing personalized newsfeeds via analysis of information novelty. In *Proceedings of the 13th International World Wide Web Conference (WWW 2004)*, pages 482–490, 2004.

[8] G. Golovchinsky. Queries? Links? Is there a difference? In *Proceedings of the 1997 Conference on Human Factors in Computing Systems, CHI 1997*, pages 407–414, 1997.

[9] V. Hatzivassiloglou, L. Gravano, and A. Maganti. An investigation of linguistic features and clustering algorithms for topical document clustering. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2001*, pages 224–231, 2001.

[10] M. Hearst. Clustering versus faceted categories for information exploration. *Communications of the ACM*, 49(4):59–61, Apr. 2006.

[11] M. A. Hearst and J. O. Pedersen. Rexamining the cluster hypothesis: Scatter/Gather on retrieval results. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'96*, pages 76–84, 1996.

[12] J. Kominek and R. Kazman. Accessing multimedia through concept clustering. In *Proceedings of the 1997 Conference on Human Factors in Computing Systems, CHI 1997*, pages 19–26, 1997.

[13] K. La Barre. Adventures in faceted classification: A brave new world or a world of confusion? In *8th International Conference of the International Society for Knowledge Organization (ISKO 2004)*, 2004.

[14] D. J. Lawrie and W. B. Croft. Discovering and comparing hierarchies. In *Recherche d'Information Assistée par Ordinateur (RIAO 2000)*, pages 314–330, 2000.

[15] L. Lee. Measures of distributional similarity. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics (ACL 1999)*, 1999.

[16] C.-Y. Lin and E. H. Hovy. The automated acquisition of topic signatures for text summarization. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*, pages 495–501, 2000.

[17] G. Marchionini and G. Geisler. The open video digital library. *D-Lib Magazine*, 8(12), Dec. 2002.

[18] K. McKeown, R. Barzilay, J. Chen, D. K. Elson, D. K. Evans, J. Klavans, A. Nenkova, B. Schiffman, and S. Sigelman. Columbia's Newsblaster: New features and future directions. In *Proceedings of HLT-NAACL 2003, Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, 2003.

[19] M. Meila and D. Heckerman. An experimental comparison of several clustering and initialization methods. *Machine Learning*, 42(1/2):9–29, 2001.

[20] N. E. Miller, P. C. Wong, M. Brewster, and H. Foote. Topic islands: A wavelet-based text visualization system. In *Proceedings of the conference on Visualization (VIS'98)*, pages 189–196, 1998.

[21] C. G. Nevill-Manning, I. H. Witten, and G. W. Paynter. Lexically-generated subject hierarchies for browsing large collections. *International Journal on Digital Libraries*, 2(2-3):111–123, 1999.

[22] G. W. Paynter and I. H. Witten. A combined phrase and thesaurus browser for large document collections. In *Research and Advanced Technology for Digital Libraries, 5th European Conference (ECDL 2001)*, pages 25–36, 2001.

[23] G. W. Paynter, I. H. Witten, S. J. Cunningham, and G. Buchanan. Scalable browsing for large collections: A case study. In *Proceedings of the Fifth ACM Conference on Digital Libraries (DL 2000)*, pages 215–223, 2000.

[24] A. S. Pollitt. The key role of classification and indexing in view-based searching. In *Proceedings of the 63rd International Federation of Library Associations and Institutions General Conference (IFLA'97)*, 1997.

[25] K. A. Ross and A. Janevski. Querying faceted databases. In *Proceedings of the Second Workshop on Semantic Web and Databases*, 2004.

[26] M. Sahami and T. D. Heilman. A web-based kernel function for measuring the similarity of short text snippets. In *Proceedings of the 15th International World Wide Web Conference (WWW 2006)*, pages 377–386, 2006.

[27] M. Sanderson and W. B. Croft. Deriving concept hierarchies from text. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'99*, pages 206–213, 1999.

[28] E. Stoica and M. A. Hearst. Nearly-automated metadata hierarchy creation. In *HLT-NAACL 2004: Short Papers*, pages 117–120, 2004.

[29] K.-P. Yee, K. Swearingen, K. Li, and M. A. Hearst. Faceted metadata for image search and browsing. In *Proceedings of the 2003 Conference on Human Factors in Computing Systems, CHI 2003*, pages 401–408, 2003.

[30] H.-J. Zeng, Q.-C. He, Z. Chen, W.-Y. Ma, and J. Ma. Learning to cluster web search results. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2004*, pages 210–217, 2004.