The Semantic GrowBag Demonstrator for Automatically Organizing Topic Facets

Jörg Diederich
L3S Research Center,
Hanover, Germany
diederich@l3s.de

Uwe Thaden
L3S Research Center,
Hanover, Germany
thaden@l3s.de

Wolf-Tilo Balke
L3S Research Center,
Hanover, Germany
balke@l3s.de

ABSTRACT

The faceted search paradigm allows to intuitively categorize the search results along orthogonal facets. Such facets, for example, can group the results along authors, publication year, or topics in case of digital libraries. One open issue, however, is how facets themselves can be organized and presented, especially if a facet is highly dynamic and is too large to be presented at once. In this paper we demonstrate the Semantic GrowBag approach to automatically organize facets, namely a topic facet, for community-specific document collections. The approach uses the documents' metadata tags to extract information about prevalent topics and then applies PageRank to determine intrinsic relations between these topics as supported by a collection's documents. We discuss a short use case for topics extracted for computer science documents in the DBLP collection and show that we can even determine a topic's development over time to enable a customizable organization of the topic facet. Given that facets have to be individually adapted to different users, user groups, or communities our algorithm is an important step towards automatic facet organization.

Categories and Subject Descriptors

H3.3.3 [**Information Search and Retrieval**]: [Clustering][Selection process]; H3.3.5 [**Online Information Services**]: Web-based services

Keywords

faceted search, tagging, topic hierarchies, Semantic Web

1. INTRODUCTION

Faceted search for document or media retrieval [3, 8] has been shown to be a valuable paradigm for exploring information spaces in a cooperative way. Facets offer different (usually orthogonal) dimensions that a user can use for document browsing. Such facets can be seen as a way to categorize content or document collections for intuitive user interaction. Given the increasing volume of today's document collections and digital libraries, such user support has been recognized as an important feature [5].

One open question when dealing with facets is in what way to organize each facet. A good design of orthogonal facets for document collections generally requires manual work, which is, however, limited due to the limited number of facets that can be presented on the search engine interface. In contrast, there can be a high number of components within each facet, which strongly requires an automatic support for the organization of the components of facets (e.g., for clustering publication years to time spans for a facet 'publication year' to avoid non-empty facet components). This has led to a rich line of extensive research in result clustering or sampling [2, 9, 1].

One of the main problems in document retrieval is that all information about a document's topics (and related information or categorizations) has to be expensively derived directly from the full texts, e.g., by using language models [4]. And what is more, the derived organization of such topic facets can differ with the views certain users or communities take with respect to a topic and may even evolve over time. Generally speaking, suitable metadata can already provide a good impression about the most important aspects of a document and only recently the advent of so-called social tagging has led to vast community-based metadata creation for documents collections (e.g., del.icio.us[1]) and media archives (e.g., flickr[2]).

In this paper we discuss how to efficiently organize one particular facet, the topic facet, using such metadata with respect to user-provided keywords. The main difference to existing (static) facet organizations is that this topic facet is sensitive with respect to time and user community. We base our work on exploiting the currently available metadata from community-specific document corpora like the DBLP collection[3] for computer science or the Medline Database[4] for medical science, where the topic-related tags and the respective dates of publications provide a strong insight into what has been considered as relevant for a community during a certain time span.

In the remainder of this paper we will explain our algorithm (whose technical properties are discussed elsewhere) in a practical use case related to the Information Retrieval community to demonstrate the feasibility of our approach.

[1]http://del.icio.us/
[2]http://www.flickr.com/
[3]http://dblp.uni-trier.de/
[4]http://medline.com/

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
Faceted Search '06, Seattle, WA, USA
Copyright 2006 ACM X-XXXXX-XX-X/XX/XX ...$5.00.

2. USING THE SEMANTIC GROWBAG

This section explains the Semantic GrowBag algorithm along a practical use case. After describing the underlying data set, all parts of the algorithm are described in detail to automatically construct a facet for related topics for the sample start tag 'Information Retrieval'. Then we go on to show how such a topic facet can change over time.

2.1 Tags in DBLP: The Data Set

DBLP is a large collection of computer science related publications with approximately 650,000 documents (as of July 2005). It also provides URLs to the electronic editions for about 330,000 publications. We were able to use 220.000 of these URLs to extract all manually annotated keywords (tags) using a wrapper-based approach [6] with manually created wrappers for three major sites offering the corresponding metadata. Only 56,000 documents of these (i.e., only about a tenth of the DBLP collection) provided proper tags and we could finally extract about 300,000 tags, resulting in a list of 130,000 unique tags.

We post-processed all tags using acronym replacement (e.g., WWW → World Wide Web) and Porter stemming [7] and filtered those tags, which are mentioned less than five times. The resulting list of 'main' tags comprises only about 8,600 tags, nevertheless representing 60% of all occurring tags. Since DBLP also provides the publication year of each document, we connected the main tags of each document with the respective publication date to show how the related tags evolve over time.

The Semantic GrowBag algorithm thus uses only a relation DBLP-ID ↔ tag-ID (about 180,000 entries) and a relation DBLP-ID ↔ year for the 53,000 documents tagged with one of the 8,600 main tags. These two relations define our 'Tagged DBLP' data set. We currently do not make use of any other available data, though we plan to examine titles, authors, citations etc. in future versions.

2.2 Algorithm Overview

The objective of the Semantic GrowBag algorithm is to construct a topic facet, i.e., a directed graph with the most closely related tags for a given start tag. Such a graph can be used, for example, to provide a faceted views in applications like document browsing or searching. As input the algorithm requires a set of documents or media objects that have been tagged with freely chosen topics[5].

The three main parts of the algorithm are as follows:

1. Determine the most closely related (top-X) tags to a start tag T and compute a biased PageRank using the top-X tags of T to include the latent relationships to other tags.

2. Derive the relations between T and its top-X tags from the biased rankings.

3. Combine the relations and the top-X tags into a single graph, comprising the topic facet.

2.3 Building Basic Graphs

This use case explains how the Semantic GrowBag algorithm creates a graph for the tag 'Information Retrieval' (IR) for all documents in our Tagged-DBLP set for the period 2000–2004. We assume that all tags have been preprocessed as described in Section 2.1. We also assume that a co-occurrence matrix for all tags has been built, to which PageRank will be applied. The value in one matrix element $M_{i,j}$ denotes how often tag i co-occurs with tag j in one document (the 'TF ' of co-occurrence), weighted with the 'IDF ' of the co-occurrence of tag i, i.e., the logarithm of the total number of tags divided by the number of tags that co-occur with tag i.

2.3.1 Part I: Top-X and Biased PageRank

At first, the 'IR' row of matrix M is extracted with the co-occurring tags of 'IR' and is sorted according to the TFxIDF values. The resulting list (referred to as the *TFxIDF list* for 'IR') is shown in table 2.3.1.

Rank	Tag	TFxIDF	TF
1	IR	669.0	200
2	Search Engine	67.0	15
3	Language Model	60.2	12
4	WWW	51.1	16
5	Web search	41.4	8
6	Query Expansion	39.9	7
7	Text mining	36.2	8
8	Indexing	29.1	7
9	NLP	28.3	6
10	Question Answer	27.0	5
...	...	...	...

Table 1: Top-10 tags from co-occurrence analysis

In the Semantic GrowBag algorithm, the top-X tags are chosen as those top tags of the TFxIDF list, which accumulate 20% of the sum of the TFxIDF values of all tags in that list (i.e., the 'mass' of the distribution), however, excluding the start tag 'IR' from the sum. In the above example, the first 22 terms would be relevant (i.e., $X = 22$), but we restrict our use case here to the top-10 for space reasons.

Afterwards, the algorithm uses PageRank (biasing to 100% on the previously determined top-X tags of 'IR') to find a new ranking. This *PageRank list* includes also latent relations among the tags, which might not be reflected in the original TFxIDF list. The main goal is to see if the start tag 'IR' itself remains the 'top tag' or if other tags are more relevant in the domain. Please note that in the TFxIDF list, the top tag is always the start tag ('IR' in our example) and no latent relations can be seen.

As shown for our example in Table 2.3.1, 'IR' indeed remains on the top, so it is an autonomous topic, and not merely a sub-topic to some more characteristic topic (i.e., it does not generally occur in only a couple of contexts). Furthermore, the ranking is significantly different between the TFxIDF list and the PageRank list. For example, 'machine learning' is on rank 3 on the PageRank list while it was only on rank 12 on the TFxIDF list. 'Machine learning' thus seems to have a stronger impact on the topic of IR than we can see from the simple co-occurrences.

2.3.2 Part II: Relations between the top-X tags

Now that we know the most important related concepts for 'IR', the goal of the second part of the Semantic GrowBag algorithm is to make the relation between the tag 'IR' and its

[5]preferentially done manually to ensure a high quality of the tagging.

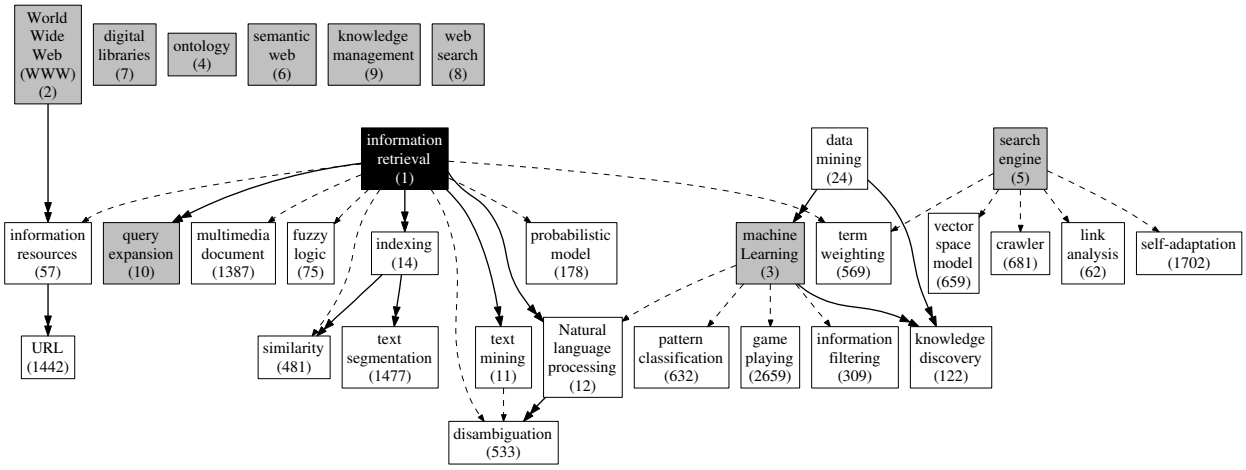

Figure 1: Topic facet for 'IR' in the period 2000–2004

Rank	Tag	Score
1	IR	148.3
2	WWW	103.8
3	Machine Learning	94.3
4	Ontology	91.9
5	Search Engine	90.4
6	Semantic Web	88.9
7	Digital Library	85.6
8	Web search	79.2
9	Knowledge Management	77.3
10	Query expansion	76.2
...	...	...

Table 2: Top-10 tags according to Biased PageRank

respective top-X tags more precise, i.e., to find super-topics and sub-topics. The basic idea is to do a pairwise comparison of the scores of two lists: the PageRank list of 'IR' and the PageRank list of its respective top-X tags ('WWW', 'machine learning', 'ontology', etc.). For the sake of brevity we will just provide two lists for 'Machine Learning' (*top*-X=14) and 'Query Expansion'(*top*-X=3) (Tables 3+4).

Rank	Tag	Score
1	Data mining	171.6
2	Machine Learning	168.1
3	IR	134.7
4	Classification	134.4
5	Neural Networks	129.7
6	Support Vector Machines	125.2
7	Decision trees	122.8
8	Pattern recognition	114.4
9	Knowledge Discovery	114.4
10	Recommender System	113.1
...	...	...

Table 3: PageRank list for 'Machine Learning'

Let us focus on what we can deduce from the list of 'query expansion'. Since the score of 'query expansion' is lower than the one of 'IR' in both PageRank lists (table 4 and tables 2.3.1), we can define 'query expansion' as a 'sub-topic'

Rank	Tag	Score
1	IR	541.3
2	Query expansion	490.5
3	Probabilistic Models	476.4
4	Search Engines	74.6
5	Web search	45.3
6	WWW	41.4
7	Data mining	37.5
...	...	...

Table 4: PageRank list for 'Query expansion'

of 'IR'. Analogously, we assume a 'super-topic' relation if the score would have been higher in both lists (which is not true for 'IR' in this use case, but e.g., for 'IR' and 'Query expansion' when using 'Query expansion' as start tag).

The explicit distinction between 'super' and 'sub' relations is useful for weighing the edges later. The comparison shown here is performed between all possible pairs of top-X tags to check for further 'sub-topic' and 'super-topic' relations.

2.3.3 Part III: Combining the relations into a graph

To show the topic facet we depict all sub-/super-topic relations in the form of a graph. The resulting graph for our sample tag 'Information retrieval' is given in Fig. 1 (for a presentational purposes we left out a rather large number of sub-topics of 'WWW' that were not connected to 'IR' and completely omitted all subtopics of related topics, if they were not connected to 'IR' in any way).

The first step in creating such graphs starts with including all the top-X tags in a set N of nodes, i.e., the start tag 'IR' itself (always depicted on a black background), and the tags 'WWW', 'Machine Learning' etc. (depicted with a grey background) as the 'seeds' for N. Then the Semantic GrowBag algorithm grows this set by collecting all sub-topics of the tags in N (depicted with a white background), i.e., those tags which are either in a sub-topic relation or in a 'super-topic' relation to some tag in N. 'Growing' N is repeated recursively until all child tags of the top-X tags have been found. Then, also all immediate super-topics of the top-X tags are added to the set N. This is to put the top-X tags into their immediate upper context and allow for query re-

laxation if it should prove necessary during browsing. In our example, this leads to the inclusion of 'data mining' as a more general parent of 'machine Learning'. We also characterize each tag in N by its rank in the PageRank list of the start tag 'IR' thus expressing how 'closely' related the tag is to 'IR'. As an example, 'text mining' (rank 11) and 'indexing' (rank 14) have a closer relation to the general concept of 'IR' than the specific concepts of 'disambiguation' (rank 533) or 'term weighting' (rank 569).

In the second step to create the graph, we visualize all 'sub-topic' and 'super-topic' relations, in which both parameters are tags from N. Edges always point from super-topics to sub-topics. But in contrast to manual facet creations, our Semantic GrowBag algorithm also provides a 'weight' for each edge as a hint for the 'confidence' in the relation. These weights depend on the support of building the relation provided by the underlying data. Weights are visualized as dashed lines (weight=1), and bold lines with two arrows (weight=2) and each weight and the direction of each edge are set as follows:

- If tag t_2 is a sub-topic of t_1, the direction is $t_1 \rightarrow t_2$ and the weight is one.
- If t_2 is a sub-topic of t_1 and t_1 is also an explicit super-topic of t_2, then also $t_1 \rightarrow t_2$, but the weight is two.

In the latter case, we use a higher weight as both t_1 and t_2 are among the top-X tags of t_1 and t_2 so both tags seem to be very important for each other. In our example, this is true for 'query expansion' and 'IR'.

If only the 'sub' relation is found, (as, for example, 'fuzzy logic' is a sub-topic of 'IR', but 'fuzzy logic' is not among the top-X tags of 'IR'), the relation is less strong (only unidirectional). Hence, we assign a lower weight.

The case where t_1 is a super-topic of t_2 (i.e., both are in the top-X of t_1), but t_2 is not a sub-topic of t_1 can only occur if t_2 is not among its own top-X tags. This is almost impossible in our case of using a biased PageRank with a 100% weight on the biasing set (we measured a probability of $< 0.1\%$ for this).

2.4 Graph Development over Time

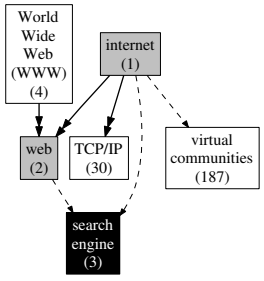

Figure 2: Topic Facet 'Search Engine' (1998-1999)

Since we additionally know the publication year of each document, our algorithm can also show the development of a topic facet for a certain tag over time. While Figure 1 depicted the complete topic facet for 'IR' using all documents in the period 2000–2004, some tags show a different development over time, i.e., the way the respective topic was perceived at different times. As an example, we show

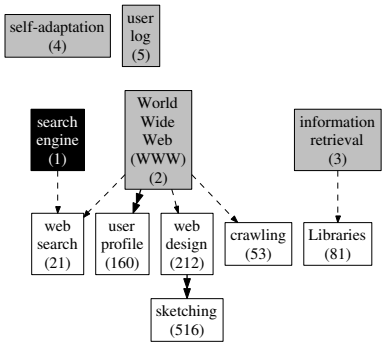

Figure 3: Topic Facet 'Search Engine' (2001-2002)

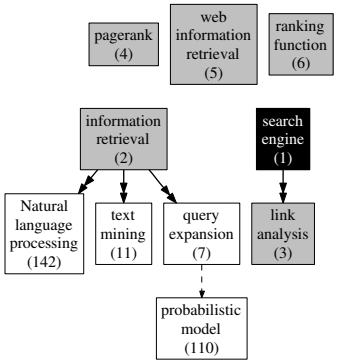

Figure 4: Topic Facet 'Search Engine' (2003-2004)

the development from 1998–2004 of the tag 'search engine'. Fig. 2 shows the topic facet of 'search engine' restricted to documents from the years 1998 and 1999, Fig. 3 for the years 2001 and 2002, and Fig. 4 for the years 2003 and 2004[6].

In 1998–1999, 'search engine' was a sub-topic of 'internet' and 'web'. In 2001–2002, search engines evolved into a research topic on its own and the relation to the Web as a super-topic was getting weaker. We can see that the tag is no longer subordinated under anything, but it shares a 'weak' connection with 'WWW' to the new topic of 'Web search'. In 2003–2004, obviously link analysis became a main topic of web search, which is also indicated by the other *top-X* topics like for instance 'pagerank'. Similarly, 'IR' has some very strong new connections in 2003–2004: 'NLP', 'text mining', and 'query expansion', which reflects very well a relevant part of the related work in 'IR' at that time.

3. THE DEMONSTRATOR

We computed the graphs for all tags for several different time periods to do a detailed evaluation of our scheme (which will be published in a more comprehensive version). To make these graphs accessible, we provided a simple web interface at http://www.l3s.de/~diederich/GrowBag/ (cf. Fig. 5). After having selected on of the available periods (bi-annual ones from 1995–2004, one 4-yearly from 2001-2004, and two 5-yearly periods for 1995–1999 and 2000–2004), the demonstrator shows a list of those graphs, where we have at least one 'strong' edge with a weight=2 (the other graphs

[6]We always aggregated the data over two years, since the data showed to be too sparse using only individual years.

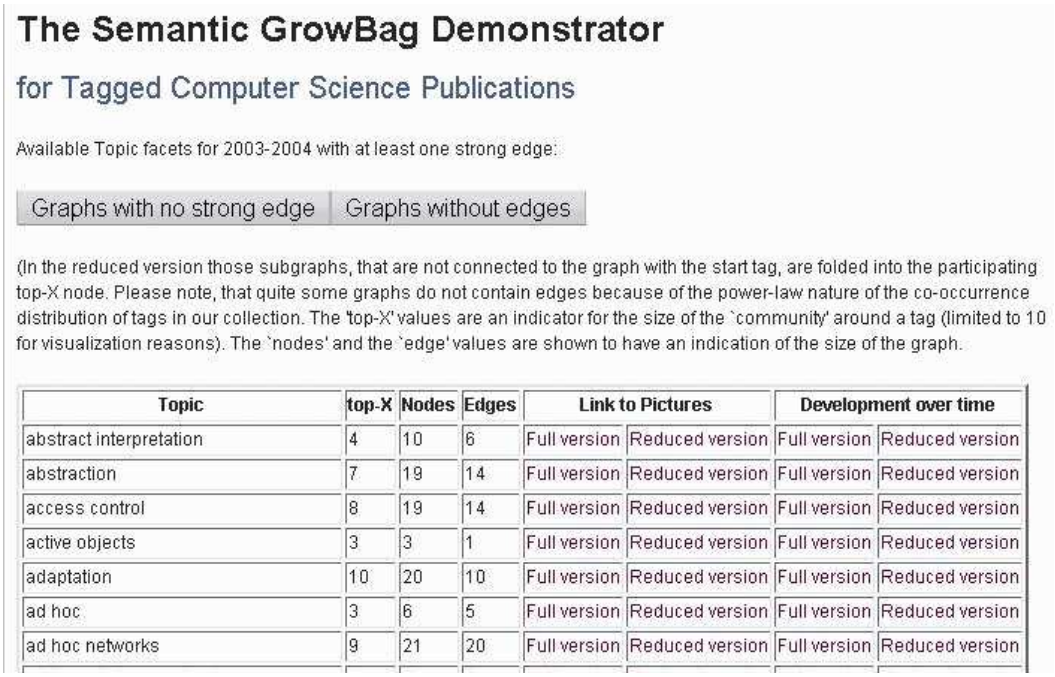

Figure 5: The Semantic GrowBag Demonstrator

can be accessed using the 'Graphs with no strong edge' button or the 'Graphs without edges' button). For each tag/topic the table shows information about the number of top-X nodes and the overall number of nodes and edges in the graph, which is intended to alleviate the search for larger graphs (due to the power-law distribution of the co-occurrences of tags, we have many graphs with no or only very few edges). In columns 5 and 6, two versions of the topic facet graph can be selected: A full version, where the topic facet is shown for all top-X tags, and a reduced version, where all those subgraphs are folded into their seeding top-X tag, that are not connected to the start tag of the table. The last two columns comprise links to special pages, which show all graphs of the selected tag for the different time periods. This is intended to better see how the selected tag has evolved over time.

4. CONCLUSIONS

In this paper we gave some insights into how to automatically organize a topic facet, which is based on metadata tags in community-specific document collections. We provided a use case that shows how our Semantic GrowBag approach uses tags and creates semantic links between them. The resulting graphs can be seen as dynamically created organization of a community-centered topic facet. This organization reflects the perception of related concepts within a certain community and for a certain time span. The organization is supported by the respective underlying document corpus and our approach can even give a (simple) measure for the confidence in the created relations. In contrast to a static organization of a topic facet, our approach results in a better distribution of the documents among the sub-topics.

In the future, we want to improve the quality of the topic facet, e.g., using a more sophisticated detection of the *top-X* tags. Furthermore, we envision a change detection scheme to automatically detect tags in a highly changing environment. We also plan to try data sets from different communities, such as the Medline Database. Finally, we want to support conjunctions of keywords as start tag to allow the disambiguation of keywords used in multiple communities.

5. REFERENCES

[1] A. Anagnostopoulos, A. Z. Broder, and D. Carmel. Sampling search-engine results. In *International World Wide Web conference*, pages 245–256, 2005.

[2] 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 *ACM SIGIR conference*, pages 318–329, 1992.

[3] M. A. Hearst. Clustering versus faceted categories for information exploration. *Commun. ACM*, 49(4):59–61, 2006.

[4] M. A. Hearst and C. Plaunt. Subtopic structuring for full-length document access. In *International ACM SIGIR conference*, pages 59–68, 1993.

[5] M. Kaeki. Findex: search result categories help users when document ranking fails. In *SIGCHI conference*, pages 131–140, 2005.

[6] A. H. F. Laender, B. A. Ribeiro-Neto, A. S. da Silva, and J. S. Teixeira. A brief survey of web data extraction tools. *SIGMOD Rec.*, 31(2):84–93, 2002.

[7] M. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[8] K.-P. Yee, K. Swearingen, K. Li, and M. Hearst. Faceted metadata for image search and browsing. In *SIGCHI conference*, pages 401–408, 2003.

[9] O. Zamir and O. Etzioni. Grouper: a dynamic clustering interface to web search results. In *International World Wide Web conference*, pages 1361–1374, 1999.