

# Automatically Created Concept Graphs using Descriptive Keywords in the Medical Domain

J. Diederich<sup>1</sup>, W.-T. Balke<sup>1</sup>

<sup>1</sup>L3S Research Center, Leibniz University of Hannover, Hannover, Germany

{diederich, balke}@L3S.de

## Summary

**Objectives:** Besides keyword search, navigational search is an important means to find relevant information in digital object collections. Such navigation is often supported by categorization systems or thesauri, which provide a hierarchical view on a particular domain and allow for browsing digital collections. Existing categorization systems, however, require large and expensive efforts for the manual creation and maintenance. Our Semantic GrowBag algorithm fully automatically creates concept graphs, i.e. directed graphs similar to categorization systems but without strong subsumption semantics. This article sketches our algorithm and evaluates it for the medical domain.

**Methods:** Our Semantic GrowBag algorithm uses descriptive keywords and exploits higher-order co-occurrences between them to create concept graphs (so-called *GrowBag graphs*) from annotated object collections. In this study, we have automatically created more than 2000 GrowBag graphs based on the Medline data set to show the applicability of our algorithm in the medical domain. For the evaluation, we first compared our algorithm to a baseline algorithm that does not take higher-order co-occurrences into account, and then compared the resulting GrowBag graphs systematically against the manually crafted MeSH thesaurus.

**Results:** Our experiments revealed that the Semantic GrowBag approach essentially increases the number of relevant relationships in comparison to a baseline approach by about 50%. Furthermore, the identified relations usually correspond to and hardly ever contradict to relationships as stated by MeSH.

**Conclusions:** The Semantic GrowBag algorithm allows creating concept graphs fully automatically. While it does not systematically exploit specifics of a domain (such as the fundamental separation between ‘drugs’ and ‘therapy’ in MeSH), the resulting GrowBag graphs are nevertheless well-suited to support navigation in digital object collections. Moreover, they can also be used to help maintaining existing categorization systems based on the actual usage of categories.

## Keywords

navigational search, subject indexing, concept graphs

In: *Methods of Information in Medicine (METHODS)*, Vol. 47(3), Schattauer, 2008.

Contact author: Wolf-Tilo Balke,  
L3S Research Center, Appelstr. 9a, 30167 Hannover, Germany  
Tel. +49 (511) 762-17712, Fax. +49 (511) 762-17779  
E-Mail: balke@l3s.de

## 1. INTRODUCTION

With the increasing size and heterogeneity of Digital Libraries (containing e.g., documents, media files, or images) metadata in form of descriptive terms has to be used to describe and summarize the objects. These can be either freely specified (often referred to as keywords or author tags), but can also be derived from controlled vocabularies, e.g. by the publisher. This can also greatly improve navigation through, searching in, and filtering of large collections [17], [10]. A large-scale generation of such keywords can sometimes even be done automatically based on the underlying objects, but this is limited mainly to collections of textual objects and is often subject to trade-offs regarding the overall quality of the keywords. Despite the problem of high costs, a large amount of manual annotations with descriptive keywords is available for many collections of digital objects. While such annotations can indeed improve the searching and filtering of object collections, their full potential for navigational search is not yet explored. Current approaches either use “related keywords” for navigation or allow navigation based on categorization systems (cf. Fig. 1) or thesauri, such as the MeSH thesaurus [20]. However, such graphs are in most cases created and maintained manually with very high efforts and are often only available for specific domains (e.g., digital libraries for biomedical informatics, cf. [5]).

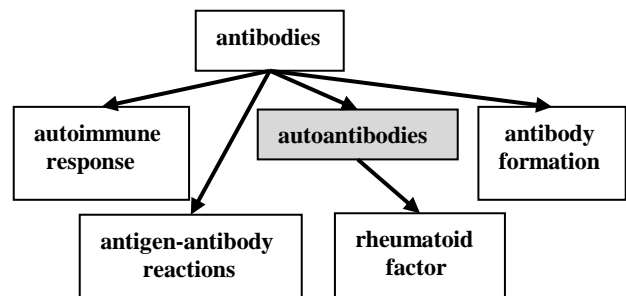


Fig. 1: Excerpt from the GrowBag graph for the keyword ‘autoantibodies’ (based on Medline)

While it is still very difficult to efficiently create high-quality categorization systems automatically, there are promising approaches to automatically create *concept graphs*. Such concept graphs differ from categorization systems only in the semantic of the relationship: While in manually created categorization systems relationships strictly denote a ‘subsumption’ relation, in concept graphs they can at best be described as ‘related to, but more specific’. Hence, concept graphs are a generalization of categorization systems where some relations may still have a ‘subsumption’ semantics.

Concept graphs have initially been designed to support navigational search also for domains, where a manual creation of categorization systems is not feasible either due to too high costs or too high dynamics. But what is more important, they can also assist in the maintenance of already existing categorization systems (such as MeSH for the biomedical domain).

Approaches to automatically create concept graphs typically exploit co-occurrences between manually specified keywords and have shown a nice performance for navigational purposes [15] when applied to automatically annotated document collections. By exploiting higher-order co-occurrences as known from computational linguistics [16], our *Semantic GrowBag* algorithm includes all transitive information about keyword co-occurrences and can be applied even on manually annotated collections with only few annotations per object. In a nutshell, it creates a complex network of keywords influencing each other, not unlike the influences between Web pages as given by links. This is reflected by the algorithm using a Biased Page-Rank algorithm [12] to find all (hidden) relationships between keywords based on higher-order co-occurrences. The details of the algorithm are given in [8] and its applicability for document retrieval in the area of computer science has successfully been demonstrated in [9], [21].

In section 2, we will briefly review existing work in creating categorization systems for typical search tasks like navigational search, query expansion or query relaxation. Section 3 will demonstrate the use of the Semantic GrowBag approach given in [8] on the Medline digital library [19], which is a bibliographic database with about 13 Million references to journal articles in life sciences, focusing on biomedicine. The evaluation is contained in section 4, followed by the conclusions in section 5.

## 2. Related Work: Categorization Systems

Categorization systems used to structure digital collections are typically acyclic directed graphs [15]. Most categorization systems (also known as classification systems or thesauri) are created manually and are strictly hierarchical. Good examples are e.g., the Dewey Decimal Classification System, the ACM Computing Classification System, the categorization system used in the Open Directory project, or the MeSH thesaurus. However, especially if such categorization systems are not maintained and updated properly, their utility can decrease rapidly (cf. the ACM classification system, which was updated last in 1998). Thus, while manually created and well-maintained categorization systems are definitely extremely valuable, they suffer from the problem that a large manual effort (and thus costs) is required for the creation and maintenance of such systems. Hence, an automatic creation is desirable. It can be based either on the objects themselves (i.e., the full text of textual objects or text segments [1]) or exploiting annotations in form of descriptive keywords (e.g., freely specified by the authors or from a controlled vocabulary) only.

Full-text approaches include those that (to some degree) create ontologies automatically, see for instance [4], [12], [3]. Several approaches exist that mostly rely on (supervised) learning techniques based on natural language processing, e.g., using language models or syntactic contexts. These approaches try to find out synonyms, sub-/superclass hierarchies, etc. by relying on the sentence structure where phrases like ‘such as...’ or ‘like e.g.,...’ imply a certain hierarchy between terms usually derived from full texts. Moreover, the belief in the correctness of derived classes

and/or hierarchies can be supported by comparison to general ontologies like WordNet or counting co-occurrences e.g., in documents retrieved by Google.

On the other hand creating categorization systems automatically based on (automatically extracted) keywords only was first proposed by Sanderson and Croft [15]. They define keyword  $X$  to subsume keyword  $Y$  if at least 80% of the documents in which  $Y$  occurs, form a subset of the documents in which  $X$  occurs, and if  $X$  is used more frequently as annotation than  $Y$ .

### 2.1 Semantics of Relations in Concept Graphs

Generally speaking, the semantics of hierarchical relationships in such usage-based concept graphs (often described as semantic maps or topic maps) is somewhat less strict compared to most manually created systems (i.e., usually not subsuming). They can best be described as meaning ‘related to, but more specific’. This ‘shallow’ semantics of the relationships stems from the fact that the relationships are based on the actual *usage* of the keywords, which might be different from abstract conceptual models. For example, we found from the Medline corpus that the keyword ‘Latvia’ is subsumed by ‘Estonia’, even though Latvia is actually not a part of Estonia. This is because all publications in Medline about Latvia were also annotated with Estonia, but the keyword ‘Estonia’ was also often used on its own.

In any case, the strict semantics of manually crafted categorization systems is often not required, e.g., for navigational search in object collections: the advantages of navigating using an inexpensive, automatically created concept graph by far outweigh its somewhat shallow semantics. However, a simple subsumption definition like in [15] does not include relationships based on higher-order co-occurrences between keywords. This might not be problematic in scenarios, where keywords are automatically extracted from full texts, so that sufficiently many keywords are available for each object. But whenever keywords are created manually, there will typically be only few keywords for each object. Thus, chances for keywords to co-occur are lower and relationships based on higher-order co-occurrences are especially valuable for creating concept graphs.

## 3. The Semantic GrowBag Approach: The Medline Use-Case

The basic idea of the Semantic GrowBag algorithm is to create concept graphs from a corpus of objects annotated with descriptive keywords including hidden relationships between individual keywords. This is done by exploiting *higher-order co-occurrences*, as known from computational linguistics [16]. As shown in Fig. 2, keywords X and Z are associated with one object (C), which is a first order co-occurrence or simply co-occurrence.

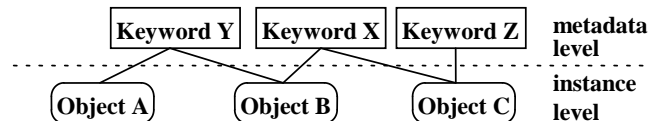


Figure 2: Example Higher-Order Co-occurrence

Keywords Y and Z are not associated with the same object, but there may still be a related keyword (X) which is associated with objects (B and C) that in turn are annotated by both keywords Y

and  $Z$ , respectively. Such higher-order co-occurrences occur more often than first-order occurrences alone and, hence, reduce the sparsity of the co-occurrence dataset. They have therefore been found to be more robust than the first-order co-occurrences improving for example word sense disambiguation algorithms [16]. Including higher-order co-occurrences has two main effects: finding additional ('hidden') relationships between keywords which cannot be found using first-order co-occurrences alone, and changing the 'strength' (i.e., the number) of existing first-order co-occurrences to include the values of higher-order co-occurrences. For extracting the higher-order co-occurrences our GrowBag algorithm uses a Biased PageRank algorithm [14]. Because the properties of PageRank are pretty well understood, it can be computed very efficiently and it also converges to a stable solution for appropriate input data.

The Semantic GrowBag algorithm comprises three main parts:

1. Compute a new co-occurrence metric based on higher-order co-occurrences
2. Find relationships between keywords, based on this metric
3. Construct for each keyword  $i$  a GrowBag graph to present a limited view on the 'neighborhood' of  $i$  (i.e., closely related (non-hierarchically subsumed) keywords + subsumed<sup>1</sup> keywords).

In step 1, we first compute a weighted co-occurrence  $m(j, i)$  between two keywords  $i$  and  $j$  as follows:

$$m(j,i) = cooc(i,j) * ICF(i) / \sum cooc(i, j) * ICF(i)$$

with  $ICF$  being used to reduce the impact of often occurring keywords and being defined as:

$$ICF = \log\left(\frac{\text{Overall number of keywords}}{\text{Total num. of keywords cooccurring with } i}\right)$$

Using the matrix  $M=(m_{ij})$  we then compute the Biased PageRank vector [14] for all keywords to include higher-order co-occurrences. To determine the number of keywords in the biasing set, we introduce the characteristic value  $P_C$  here, typically 15%, that essentially defines the size of the set of closely related neighbors for each keyword and limits the amount of transitivity of co-occurrence. It is the only 'tuning knob' of our scheme and its value depends on the underlying object collection. In step 2 we basically compare the PageRank vectors for any pair of keywords  $(i, j)$ : If one keyword  $i$  achieves a higher score in both vectors, it is defined to subsume the other keyword; if both keywords are closely related in both vectors, we define the confidence in the subsumption relation as 'high' and simply name the subsumption relation to be 'strong'. Finally we create a GrowBag graph for each keyword starting from the closely related neighbors and including all subsumed keywords. For the exact details of the algorithm the reader is referred to [8].

For a better understanding, we will demonstrate all parts to construct the GrowBag graph for 'natural family planning' (using all publications in Medline from 1990-2005 and the characteristic value  $P_C := 15\%$ ). The first two parts of this use case compute the relationship between two sample keywords 'natural family planning' and 'ovulation detection'.

**Finding Higher-Order Co-occurrences:** After having created a pair-wise co-occurrence matrix  $M$  including the logarithmic weights  $ICF$ , the direct neighbors of both keywords 'natural family planning' and 'ovulation detection' are computed. For 'natural family planning', table 1 shows the list of co-occurring keywords, sorted by the weighted co-occurrence.

**Table 1: Ranked Keyword List for 'natural family planning'**

Rank	Keyword	Coocc.	Weigh. Coocc.
1	natural family planning	126	461.3
2	family planning, behavioral methods	126	404.4
3	family planning	126	291.3
4	Contraception	63	152.6
5	developed countries	54	123.4
6	cervical mucus method	25	115.6
7	ovulation detection	26	115.1
8	Reproduction	41	102.7
9	sympto-thermal method	20	95.6
10	research methodology	39	89.0

The keywords 'family planning, behavioral methods' and 'family planning' always co-occur with 'natural family planning' (for all 126 occurrences of 'natural family planning' in Medline articles), i.e., papers in Medline are often not only annotated with the most specific keyword, but with a whole path of keywords (possibly inspired by the MeSH thesaurus). Weighting the co-occurrences reduces the impact of often occurring keywords significantly. 'family planning', for instance, is used 9408 times in our dataset and, thus, achieves a lower weighted co-occurrence than 'family planning, behavioral methods' occurring only 315 times. Using our characteristic value  $P_C$  of 15% only the first seven rows of the table determine the direct neighbors for 'natural family planning'. For 'ovulation detection', the sorted list of co-occurring keywords is shown in table 2.

<sup>1</sup> cf. Section 2.1 for the discussion of the semantics of 'subsume'

**Table 2: Ranked Keyword List for ‘ovulation detection’**

Rank	Keyword	Coocc.	Weigh. Coocc
1	ovulation detection	35	155.0
2	laboratory procedures	35	105.6
3	laboratory examinations and diagnoses	35	96.9
4	natural family planning	26	95.2
5	examinations and diagnoses	35	91.5
6	family planning, behavioral methods	27	86.7
7	family planning	31	71.7

Like for the previous keyword, annotation tend to contain whole paths of keywords: in this example ‘laboratory procedures’, ‘laboratory examinations and diagnoses’, and ‘examinations and diagnoses’ always co-occur with ‘ovulation method’. However, ‘natural family planning’ co-occurs only in 74% of the cases with ‘ovulation detection’ (thus, it would not be found by an approach using only basic co-occurrences if they co-occur in  $> 80\%$  of the cases [15]). This time using a characteristic value  $P_C := 15\%$  the direct neighbors are restricted to the first four keywords.

Based on the direct neighbors, now PageRank scores are computed for both keywords to finish part 1. Table 3 lists the head of the (sorted) PageRank score vector when biasing on the seven direct neighbors of ‘natural family planning’.

**Table 3: PageRank Score Vector for ‘natural family planning’**

Rank	Keyword	PageRank Score
1	family planning	703.4
2	developed countries	669.0
3	Contraception	610.5
4	family planning, behavioral methods	490.1
5	natural family planning	474.8
6	ovulation detection	437.7
7	cervical mucus method	436.3
8	Population	348.2
9	demographic factors	340.6
10	developing countries	325.0

It can be observed (a) that the ranking in Table 3 has changed compared to the ranking based on weighted co-occurrence as shown in Table 2 and (b) that four keywords managed to get a higher score than ‘natural family planning’, which makes them prime candidates to subsume ‘natural family planning’ in our approach. Similarly, table 4 shows the head of the (sorted) PageRank score vector when biasing on the four direct neighbors of ‘ovulation detection’. This time three keywords achieve higher

scores than ‘ovulation detection’, among them ‘natural family planning’ as potential candidate to subsume ‘ovulation detection’.

**Table 4: PageRank Score Vector for ‘ovulation detection’**

Rank	Keyword	PageRank Score
1	laboratory examinations and diagnoses	846.5
2	laboratory procedures	809.0
3	natural family planning	775.6
4	ovulation detection	754.2
5	developing countries	338.2
6	Population	335.8
7	demographic factors	328.7

**Finding Hidden Relationships:** To actually determine the relationship between ‘natural family planning’ and ‘ovulation detection’ we compare the scores of both keywords in both lists<sup>2</sup>. Here ‘natural family planning’ achieves a higher score than ‘ovulation detection’ in both PageRank vectors. Thus, ‘natural family planning’ is a candidate for subsuming ‘ovulation detection’. Finally we have to post-filter the candidates based on their rank in both sorted PageRank vectors shown in Table 3 and 4. Both keywords occur within the top neighbors of both lists. Hence, our algorithm defines the subsumption relationship between ‘natural family planning’ and ‘ovulation method’ as a strong relationship.

**Building the GrowBag Graphs:** After having performed the above steps for all different keywords, we can construct the actual concept graphs. In our example a GrowBag graph for ‘natural family planning’ is created (cf. Fig. 3), using the following steps: Start with the related keywords of ‘natural family planning’, i.e., ‘family planning’, ‘developed countries’, ..., ‘ovulation detection’, and ‘cervical mucus method’ (the start keyword is depicted with a black background, all related keywords with a gray one). Now add all keywords, which subsume one of the above keywords related to ‘natural family planning’ (e.g., ‘population’, ‘developing countries’), and add recursively all those keywords which are subsumed by one of the above keywords. Finally add all known subsumption relationships between the above keywords. In our depiction of the graph the bracketed numbers in the boxes below each keyword denote the rank in the sorted PageRank vector to give an indication how close the keyword is related to the start keyword. For example, ‘ovulation detection’ is on position 6 in the PageRank score vector of ‘natural family planning’ (cf. Table 3). Relationships are drawn with dashed lines for relationships with a weak confidence and with bold lines having a two-headed arrow in case of a strong confidence. All GrowBag graphs for the Medline 2006 collection can be found at [18].

<sup>2</sup> Our algorithm also applies ‘tail cutting’ to remove keywords with rather small scores (which does not impact this example).

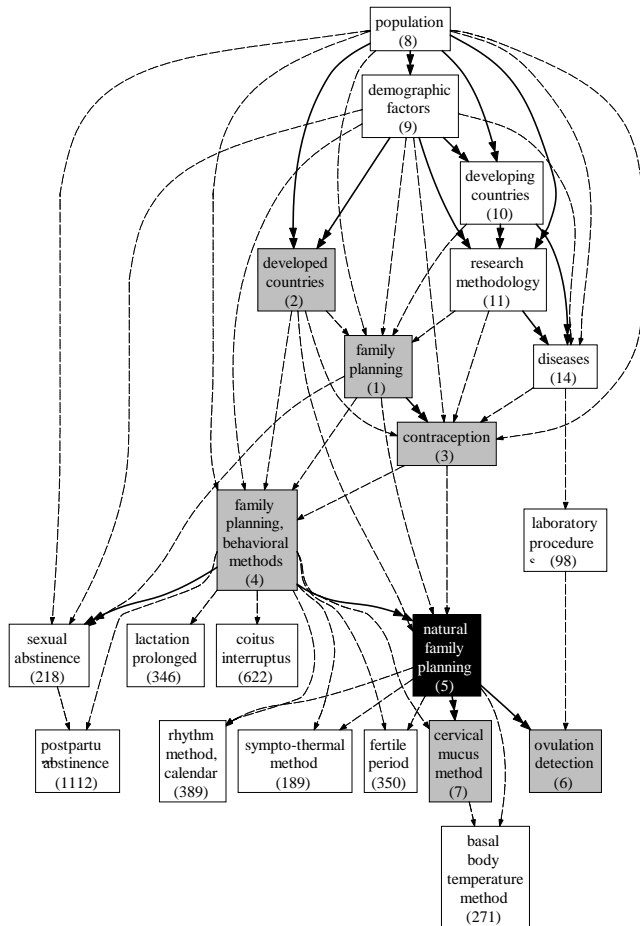


Fig. 3: GrowBag graph for ‘natural family planning’

## 4. RESULTS AND DISCUSSION

As demonstrated in the previous section, we have applied the Semantic GrowBag algorithm to the medical domain using all publications in Medline from 1990-2005 and the characteristic value  $P_C := 15\%$ . In this section all created GrowBag graphs have been evaluated according to three criteria:

- The number of relationships found by GrowBag by inclusion of the higher-order co-occurrences as compared to the baseline approach including first-order co-occurrences only.
- The stability of the included relationships regarding changes in the co-occurrence graph.
- The existence of an optimum for the control parameter  $P_C$  for the Medline dataset.

Furthermore, we also performed an initial validation of the identified relationships by comparing them to the ‘gold standard’ in biomedicine, the MeSH thesaurus. This validation is intended as a first step only, basically to find out whether the relations contradict with those in MeSH or confirm them. A user study is planned as future work to actually show the value of using GrowBag graphs. We start with a short description of the utilized dataset, a subset of the Medline database.

### 4.1 The Medline Data

Medline provides two different kinds of descriptive keywords: The MeSH classifiers of each publication (MeSH headings) as annotated by human experts at NLM and those specified by external sources: NASA, KIE (Kennedy Institute of Ethics, Georgetown University), and PIP (Population Information Program, Johns Hopkins School of Public Health) as specified in the ‘KeywordList’ tag of the Medline XML schema. In this study, we use the externally specified keywords to have a mix of freely specified keywords and keywords from controlled vocabularies (mainly the ‘PIP’ source, where 25% stem from the POPLINE thesaurus [25]). The POPLINE thesaurus overlaps to 15% with MeSH and contradicts only in one case which allows for a comparison with MeSH (cf. Section 4.4).

Table 5 shows the characteristics of the Medline data set using the externally created keywords for different subsets of publications depending on the publication year.

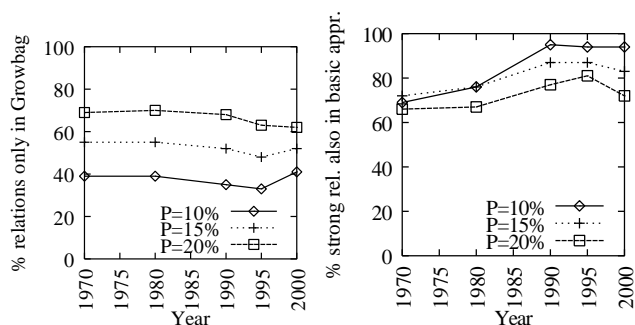
Table 5: Development of Descriptive Keywords in Medline

Period	# of keywords	# of docs	# keyw. per doc
1970-2005	1547527	145755	10.6
1980-2005	1349598	130623	10.3
1990-2005	777982	89848	8.7
1995-2005	383940	57766	6.6
2000-2005	58284	22694	2.6

As the dominating source ‘PIP’ (with on average 20 keywords per document including many ‘path’ expressions and, thus, contributing most to the generated GrowBag graphs) provides annotations only until 2000, the number of keywords per document is decreasing for the more recent periods. In most evaluations, we use the subset of publications in the period 1990-2005 for computational reasons. As we heavily rely on the co-occurrence analysis for the keywords, we have analyzed the distribution of the number of co-occurring keywords per keyword (which is the outlink distribution of the co-occurrence matrix  $M$ ) for the period 1990-2005 and found that the tail of the curve indeed is power-law distributed (other periods show similar distributions) as is also confirmed by other studies [1]. The average keyword co-occurrence is about 60 for the period 1970-2005 and decreases to 30 for the period 1995-2005 because of the diminishing percentage of the ‘PIP’ annotations.

### 4.2 Benefit of Higher-Order Co-occurrences

In this section, we will show that adding higher-order co-occurrences indeed helps to increase the number of identified relationships.



**Fig. 4: (a) Additional relationships not found by the basic approach / (b) Strong relations found by the basic approach**

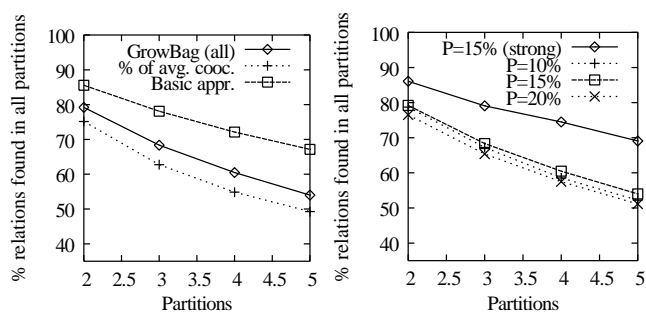
Fig. 4 (a) depicts the additional relationships found by GrowBag that are not identified by the basic approach by Croft and Sanderson. Depending on the characteristic parameter  $P_C$  and the size of the underlying data set as determined by the start year of the considered period on the x-axis, the Semantic GrowBag algorithm can identify 40-75% additional relationships.<sup>3</sup>

Fig. 4 (b) depicts the ‘strong’ relationships found by GrowBag that are also identified by the basic approach. While only 25-60% of all relationships from GrowBag are also found by the basic approach, this holds for 60-90% of the strong relationships from GrowBag. This percentage is lower for larger datasets because the larger the data set, the lower the probability that 80% of the documents annotated with one keyword are also annotated with a second keyword. In addition, we found that relationships identified by GrowBag hardly ever contradict with those found by the basic approach ( $< 0.5\%$ ). Finally, we found the distribution of the PageRank scores to be also power-law distributed. Hence, our tail cutting restricts the search for hidden relationships in the sorted PageRank score list to on average the first 600 elements (for the 1990-2005 dataset, 700 for 1980-2005). This was also independent of the chosen value for  $P_C$ .

### 4.3 Stability and Existence of an Optimal $P_C$

In this section, we examine the stability of the relationships found by the GrowBag algorithm. The main idea is to split the Medline dataset into  $n$  disjunctive and equally-sized partitions and to run the algorithms separately on each partition. Afterwards, we compare how many of the relationships found using the overall dataset (called *overall-relationships* in the following) can still be found in *all* partitions. This evaluation was done for the period 1990-2005, the results are based on a 5-fold cross-validation, i.e., the partitioning of the original dataset was done randomly five times using different seeds for the random number generator. From Fig. 5 (a), several observations can be made:

<sup>3</sup> We extended the basic approach with a similar ‘tail cutting’ as GrowBag to get a comparable quality of the relations. Hence, the increase in the number of identified relations is really due to the contribution of the higher-order co-occurrences and not an effect of using tail cutting only for our algorithm.



**Fig. 5: Regain percentage: (a) Comparison GrowBag vs. basic approach / (b) Influence of parameter  $P_C$**

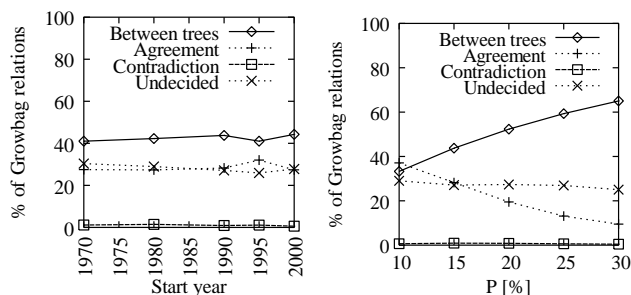
The GrowBag algorithm can regain on average 55-80% of the overall-relationships in all partitions (the minima/maxima from the 5-fold cross validation were all within  $\pm 1\%$ ). This *regain percentage* decreases with the number of partitions because smaller partitions are less connected as indicated by the curve ‘% of avg. cooc’ in Fig. 5 (a), depicting the percentage of the average number of co-occurring keywords in relationship to the average number of 46 co-occurring keywords per keyword in the full dataset. The regain percentage is about 15% lower as for the baseline approach (again extended with ‘tail cutting’ for comparison reasons). This represents the basic trade-off for finding additional hidden relationships based on higher co-occurrences, which are not found at all by the baseline approach, but which are at the same time more sensitive to a possible link removal caused by the partitioning scheme used to examine stability. We also discovered that the tail cutting, i.e. ignoring keywords with a too low Page-Rank score, contributed an increase of about 10-15% to the regain percentage in the GrowBag approach.

Besides finding additional (hidden) relationships, the second major advantage of our approach is the ability to derive optimal values for the control parameter  $P_C$ , which determines the size of the neighborhood used for the Biased PageRank computation. Fig. 5 (b) shows two main results: the results for the different values of  $P_C$  are pretty close, which underlines that GrowBag is not too sensitive against a non-optimal choice of the characteristic parameter  $P_C$  ( $P_C = 15\%$  seems to be an optimal value for the given Medline dataset) While it is possible to relax the ad-hoc 80% threshold of the basic approach [15] to find more relationships based on first-order co-occurrence only, an optimal choice cannot be made: The stability decreases monotonically with a decreasing threshold value. Figure 5 (b) additionally shows that the regain percentage of the ‘strong’ relationships for  $P_C = 15$  is higher than for all relationships and comparable to the regain percentage of the basic approach. This is an indicator that the distinction between weak and strong relationships in the Semantic GrowBag approach is reasonable and that the loss in regain percentage for our GrowBag approach is mainly affecting weak relations only.

### 4.4 The Concept Hierarchy Compared to MeSH

This last evaluation section finally compares the GrowBag concept relationships to the MeSH thesaurus, a ‘gold standard’ in the area of biomedicine which contains about 23,000 concepts orga-

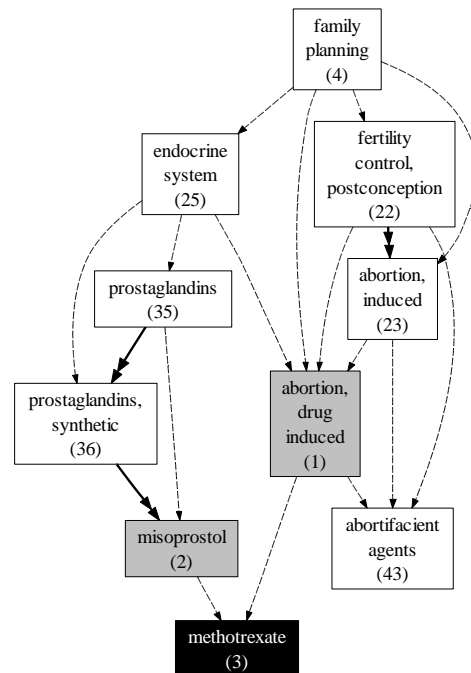
nized in 16 different trees (e.g., about anatomy, organisms, drugs, information science, or geographic locations). Since MeSH is a (quite complete) categorization system, it focuses on a specific type of relationships only, namely categorical subsumptions. Of course these special relationships (at least the part that is often used in literature) should also be reflected by our more general relationships. Therefore, the main idea is to check whether the GrowBag relationships can also be found in the MeSH thesaurus and if yes, whether the direction of the relationships is the same or contradicting MeSH.



**Fig. 6: Comparison of GrowBag and MeSH: (a) Depending on period / (b) Depending on P**

Figure 6 (a) depicts the development for  $P_C := 15\%$  for different periods. It can be seen that the rate of agreement decreases slightly, the older the dataset becomes. This can be explained with the changes in MeSH (it is updated yearly, we used the 2006 version) and the corresponding changes and developments in the area of biomedicine. Therefore, it is important to limit the data used to build a GrowBag graph to a certain period to avoid having artifacts because of the changes in the area. Figure 6 (b) depicts the results for different values of  $P_C$  for 1990–2005. Between 10–40% of the relationships involving MeSH concepts can directly be confirmed by relationships in MeSH and this percentage decreases for larger neighborhoods. For the optimal neighborhood of  $P_C := 15\%$  as determined by our stability analysis, about 30% of the GrowBag relationships involving MeSH concepts are strictly in agreement with MeSH. On the other hand contradictions between GrowBag relationships and MeSH relationships are always less than 1%. Hence, the vast majority of GrowBag relationships showing a hierarchical character are never contradicting MeSH relationships<sup>4</sup>.

<sup>4</sup> Only 34% of all Medline author keywords accounting for about 41% of all keywords can actually be matched to MeSH concepts based on string matching. This excludes about 70% of the GrowBag relationships from this comparison to MeSH (independently from the examined Medline period).



**Fig. 7: GrowBag hierarchy for the term ‘methotrexate’**

In summary, we can directly confirm about 30% of GrowBag relationships and hardly ever contradict MeSH. But let us also consider the remaining relationships between related concepts in medical research found by the Semantic GrowBag. The ‘undecided’ relationships which neither agree with nor contradict MeSH (which is the case, for instance, when GrowBag connects two keywords that are siblings in a MeSH tree) comprise about 30% of the GrowBag relationships. Moreover, between 40–70% of the relationships involve keywords from two different MeSH trees, between which there are no relationships in MeSH because the relation is not of a category subsumption type. This percentage also increases strongly with the size of the neighborhood. As an instance for such connections let us consider our running ‘family planning’ example and focus on the GrowBag tree for the drug ‘methotrexate’ (cf. Figure 7) whose pharmaceutical action is classified in MeSH under ‘abortifacient agents, nonsteroidal’. Methotrexate is correctly classified in MeSH tree D as a drug, but our GrowBag graph also clearly shows it to be an abortifacient by subsuming it hierarchically under ‘abortion, drug induced’, which in turn is a heavily related, but more specific concept of ‘family planning’. Please note that the semantic notion of the directed relation here is definitely not of a ‘subsumption’ kind, since the abortifacient characteristic is in fact an adverse effect of methotrexate, whose main application is as an antimetabolite and antifolate drug in the treatment of cancer and autoimmune diseases. But as reflected by the Medline corpus currently the abortifacient characteristic of methotrexate is heavily researched in more detail. Thus, although these are indeed sensible relationships, all three terms in the concept hierarchy belong to different trees in MeSH: whereas ‘methotrexate’ belongs to tree D, ‘abortion, drug induced’ belongs to tree E and ‘family planning’ belongs to tree N. Hence, no comparable relationships can be gained from a thesaurus like MeSH.

Besides, ‘abortion, induced’ and ‘abortion, drug induced’ are put into a correct hierarchy by the GrowBag, MeSH collects both keywords under the term ‘abortion, induced’ referring to a larger number of techniques. Looking further at the graph we also get the correct classification of induced abortion as ‘fertility control, postconception’, which is itself not a MeSH keyword, but is derived from a different controlled vocabulary (as available on POPLINE and being the basis for about 25% of the annotations provided by the external source ‘PIP’). Moreover, we can also see a clear connection with the ‘endocrine system’ (MeSH tree A). Here a correct relationship to the hormones ‘prostaglandins’ and their synthetic equivalents is made and leads to the drug ‘misoprostol’ that is classified by MeSH as ‘a synthetic analog of natural prostaglandin’. Finally, the relationship between ‘misoprostol’ and ‘methotrexate’ is given by a large number of studies using both drugs together for abortions in clinical trials (see e.g., [6], [7], [11]) where, however, methotrexate was the less common drug with respect to the application in abortions: misoprostol is approved as an abortifacient in many countries, whereas (as stated above) for methotrexate it is only an adverse effect.

As a last evaluation, we checked how those relationships which are in agreement with MeSH, are distributed among the different MeSH trees. Figure 8 depicts the distribution of all confirmed relationships for the period 1990-2005. About 30% of the identified relationships are in the last MeSH tree; tree Z named ‘Geographic Locations’. For such trees, the meaning of the hierarchical relationships is absolutely clear: For example, Florida is definitely a part of the United States, which is a part of Northern America, which in turn is a part of America. The semantics of these hierarchical relationships is less clear for other trees, where the hierarchies imposed by MeSH might even be subject to discussion among experts, as well as subject to changes over time.

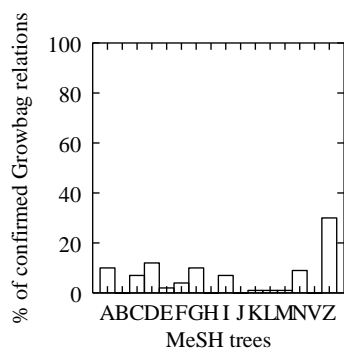


Fig. 8: Distribution of confirmed relationships for MeSH trees

#### 4.5 Further Application Examples

The GrowBag graphs shown in this study present the biomedical domain from the perspective of state-of-the-art research as represented by the studies in Medline. In this way, they can assist any person working in the medical domain in finding relevant concepts and subsequently look for further information using legacy search engines to finally participate in the medical research advances even when they are not themselves active in such research. As such, the GrowBag graphs generated from the Medline corpus can only contain relations which are already known for a while unless path expressions are used when annotating documents (which is partly the case for ‘PIP’ annotations).

Additionally, the Semantic GrowBag algorithm has the potential to help in a large variety of systems in medicine and health care. It can be used, for example, for the maintenance of ontology and terminology systems [5] (e.g., MeSH) to deal with the ever-evolving medical knowledge. In this context, it can propose new relations to be considered or to identify relations which are no longer important to assist the manual update process. Furthermore, it can complement tools for finding links between different thesauri to create and maintain meta-thesauri such as the Unified Medical Language System [13]. This is possible in case a document collection exists where documents are annotated with concepts from different thesauri, as shown in this study where 25% of the ‘PIP’ annotation actually matched MeSH categories. In this way, the integration of information currently available in separate systems (patient records, hospital information systems) into a single health information system [24] can also be supported.

Using GrowBag graphs to assist in knowledge management tasks can also be helpful in local settings, for example, modeling the knowledge of a single healthcare organization as part of a clinical decision support system [22], as long as there is an annotated set of objects which can be utilized for the modeling task. For this purpose, electronic patient records may be used which are supposed to be integrated more tightly with decision support systems in the future [23] in anyway.

## 5. CONCLUSIONS

As a result of extensive annotation efforts many objects in digital collections are associated with meaningful keywords to aid searches for these objects. Our Semantic GrowBag algorithm allows to create hierarchical relationships between such keywords to construct concept graphs without requiring additional information, e.g. from full-texts. Despite the somewhat weaker semantics of usage-based hierarchical relationships, they have proved to be very useful for navigational searches in large object collections. Our evaluation based on the Medline corpus shows that our scheme can indeed achieve about 50% additional relationships due to the introduction of higher-order co-occurrences. We then evaluated these additional relationships by a sampling method and proved that the relationships are reasonably stable regarding changes in the dataset.

Moreover, taking the well-known MeSH thesaurus as comparison, we show that our relationships are often confirmed and hardly ever contradict this hand-crafted thesaurus, but on top provide a large number of additional relationships based on how keywords are actually used for annotations. We showcased such additional relationships to be very helpful in finding relationships for instance between keywords from different MESH trees. Please note, however, that the keywords used in medical publications are, in fact, purposefully assigned for indexing by experienced and knowledgeable authors or expert indexers associated with libraries or publishers. The quality of the resulting concept graph (and thus the cost-savings in contrast to a manually designed thesaurus) may strongly differ for areas, where publications are tagged with a wide and probably less accurate variety of arbitrary keywords.

As future work, we plan to conduct a user study to actually show the benefits of our algorithm while using them, for example, for document retrieval. To achieve more matches between the descriptive keywords and the MeSH terms, we will try using the UMLS meta-thesaurus. Furthermore, we plan to extend our ap-

proach into the collaborative tagging domain that has become very popular recently due to websites like flickr, del.icio.us and the like. In this case, the different cognitive aspects of tagging such as ambiguous terms, synonyms, or basic level variations have to be taken into account [10]. While this was only partially a problem for the Medline dataset, it is a serious problem for object collections from collaborative tagging efforts (e.g. the connotea or CiteULike dataset) where several people can tag the same resource and tagging is not only used for indexing, but also, for example, for memorizing quality (e.g., ‘interesting paper’ or ‘good research’). We also want to take a closer look to developments over time in categorization systems to be able to identify emerging topics or topics, which have decreased in interest.

## REFERENCES

- [1] Begelman G, Keller P, Smadja F. Automated Tag Clustering: Improving Search and Exploration in the Tag Space. In Collaborative Web Tagging Workshop at WWW2006, Edinburgh, UK. ACM Press; 2006.
- [2] Chuang SL, Chien LF. Taxonomy Generation for Text Segments: A practical web-based approach. ACM Trans. Inf. Syst. 2005; 23(4):363–396.
- [3] Cimiano P, Handschuh S, Staab S. Towards the self-annotating web. In Int. Conf. on the World Wide Web (WWW). New York, NY, USA. ACM; 2004. p. 462–471.
- [4] Cimiano P, Völker J. Text2onto - a Framework for Ontology Learning and Data-driven Change Discovery. In Int. Conf. on Applications of Natural Language to Information Systems (NLDB). Alicante, Spain. Springer; 2005. p. 227–238.
- [5] Cimino JJ, Zhu X. The practical Impact of Ontologies on Biomedical Informatics. IMIA Yearbook of Medical Informatics. Methods Inf Med. 2006;45 Suppl 1:124–35.
- [6] Creinin MD, Vittinghoff E, Keder L, Darney PD, Tiller G. Methotrexate and misoprostol for early abortion: a multicenter trial. I. Safety and efficacy. Contraception. 1996; 53:321–327.
- [7] Creinin MD, Vittinghoff E, Schaff E, Klaisle C, Darney PD, Dean C. Medical abortion with oral methotrexate and vaginal misoprostol. Obstet Gynecol. 1997; 90:611–616.
- [8] Diederich J, Balke WT. The Semantic GrowBag Algorithm: Automatically Deriving Categorization Systems. In Proc. of Europ. Conf. on Research and Advanced Technology for Digital Libraries (ECDL), Budapest, Hungary. Springer; 2007.
- [9] Diederich J, Balke WT, Thaden U. Demonstrating the Semantic GrowBag: Automatically Creating Topic Facets for FacetedDBLP. In Proc. of ACM IEEE Joint Conf. on Digital Libraries (JCDL), Vancouver, Canada. IEEE; 2007.
- [10] Golder S, Huberman B. Usage patterns of collaborative tagging systems. J of Information Science. 2006; 32(2):198–208.
- [11] Hausknecht RU. Methotrexate and misoprostol to terminate early pregnancy. New England J Med., 1995; 333:537–540.
- [12] Hearst MA. Automatic Acquisition of Hyponyms from Large Text Corpora. In Int. Conf. on Computational Linguistics. Nantes, France. 1992. p. 539–545.
- [13] Lindberg DAB, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inf Med. 1993; 32:281–291.
- [14] Page L, Brin S, Motwani R, Winograd T. The Pagerank Citation Ranking: Bringing Order to the Web. Tech. Report, Stanford University, 1998. (available at <http://www.stanford.edu/~backrub/pageranksub.ps>, last accessed on 10.8.2007)
- [15] Sanderson M, Croft B. Deriving concept hierarchies from text. In Proc. of Int. ACM SIGIR Conf. on Research and Development in Information Retrieval. Berkeley, CA, USA. ACM; 1999. p. 206–213.
- [16] Schütze H. Automatic word sense discrimination. Comput. Linguist. 1998; 24(1):97–123.
- [17] Stuckenschmidt H, de Waard A, Bhogal R, Fluit C, Kampman A, van Buel J, et al. A topic-based browser for large online resources. In Proc. of Int. Conf. on Knowledge Engineering and Knowledge Management (EKAW), Whitebury Hall, UK. Springer; 2004. p. 433–448.
- [18] <http://demo.l3s.uni-hannover.de/Medline> (last accessed on 10.08.2007).
- [19] <http://www.nlm.nih.gov/pubs/factsheets/medline.html> (last accessed on 10.08.2007).
- [20] <http://www.nlm.nih.gov/pubs/factsheets/mesh.html> (last accessed on 10.08.2007).
- [21] <http://dblp.l3s.de> (last accessed on 10.08.2007). Updated weekly.
- [22] Peleg M, Tu S. Decision support, knowledge representation and management in medicine. Methods Inf Med. 2006;45 Suppl 1:72–80.
- [23] Jaspers MW, Knaup P, Schmidt D. The computerized patient record: where do we stand? Methods Inf Med. 2006;45 Suppl 1:29–39.
- [24] Kuhn KA, Giuse DA. From hospital information systems to health information systems. Problems, challenges, perspectives. Methods Inf. Med. 2001;40(4):275–87.
- [25] [http://db.jhuccp.org/ics-wpd/popweb/keywords/POPLINE\\_Keyword\\_Guide.pdf](http://db.jhuccp.org/ics-wpd/popweb/keywords/POPLINE_Keyword_Guide.pdf) (last accessed on 10.08.2007).