Interactive Thesaurus Assessment for Automatic Document Annotation

Kai Eckert
Computer Science Institute
University of Mannheim
A5, 6 68159 Mannheim,
Germany
kai@informatik.uni-mannheim.de

Heiner Stuckenschmidt
Computer Science Institute
University of Mannheim
A5, 6 68159 Mannheim,
Germany
heiner@informatik.uni-mannheim.de

Magnus Pfeffer
University Library
University of Mannheim
Schloss, 68159 Mannheim,
Germany
pfeffer@bib.uni-mannheim.de

ABSTRACT

The use of thesaurus-based indexing is a common approach for increasing the performance of document retrieval. With the growing amount of documents available, manual indexing is not a feasible option. Statistical methods for automated document indexing are an attractive alternative. We argue that the quality of the thesaurus used as a basis for indexing in regard to its ability to adequately cover the contents to be indexed is of crucial importance in automatic indexing because there is no human in the loop that can spot and avoid indexing errors. We propose a method for thesaurus evaluation that is based on a combination of statistical measures and appropriate visualization techniques that supports the detection of potential problems in a thesaurus. We describe this method and show its application in the context of two automatic indexing tasks. The examples show that the methods indeed eases the detection and correction of errors leading to a better indexing result. Please refer to http://www.kaiec.org for high resolution media of all figures used in this paper, as well as an animated presentation of the interactive tool.

Categories and Subject Descriptors

H.3.1 [**Content Analysis and Indexing**]: Thesauruses; H.3.7 [**Digital Libraries**]; H.5.2 [**User Interfaces**]; H.1.1 [**Systems and Information Theory**]: Information Theory

General Terms

Algorithms, Experimentation, Measurement

Keywords

Thesaurus-Based Retrieval, Suitability, Visualization

1. MOTIVATION

Advanced methods for retrieving documents based on their contents are becoming more and more important in many application areas. Today information about virtually any topic is accessible in digital form through digital libraries many of which are accessible over the web. It is common knowledge in the digital library community that semantic annotations of documents in terms of keywords from controlled vocabularies and thesauri are the key to successful search because they leverage the problems of standard methods from information retrieval. In particular, they solve the problem of using different terms to talk about the same topic. This is done by explicitly representing information about synonymous terms and relating them to a preferred term that is used to describe the content of a document. The benefits of using a thesaurus for annotating documents comes at the price of the effort needed for annotating large document sets. Traditionally, this annotation is done manually by specialists that read the abstract of a document and decide which of the preferred terms in a thesaurus best describe its content. With the increasing amount of information manual annotation is not feasible any more. Therefore, techniques for automatically annotating documents with terms from a thesaurus have received significant attention recently.

With the increased use of automatic indexing techniques a fundamental problem of thesauri is becoming more pressing: while the limitation of annotations to preferred terms in the thesaurus improves document search, the quality of the annotations significantly depend on the quality of the thesaurus. If the terms of the thesaurus do not adequately cover the topics represented in the documents, the annotation will be sub-optimal. This problem has been less pressing when annotation is done manually, because the human annotator can compensate for a sub-optimal thesaurus by choosing non-trivial combinations of existing terms or even by modifying the thesaurus to better cover a particular topic. In the presence of fully automatic annotation, the issue of thesaurus quality becomes an essential one, because existing methods for automatic annotation are not able to compensate for missing or misleading terms.

We argue that methods for automatic thesaurus-based document annotation require an additional thesaurus assessment step in which the suitability of the thesaurus as a basis for annotating a certain document set is assessed. The goal of this step is to avoid the problems mentioned above and to unveil hidden problems in the thesaurus that first need to be fixed before automatic annotation can be performed successfully. We further argue that this assessment step has to be an interactive one. There has been some work that aims at computing the quality of a thesaurus or more generally an ontology either simply on the basis of the thesaurus or based on the correspondence between the thesaurus and the corpus of documents. While such methods are a good starting point for an analysis because they can provide interesting statistics about the thesaurus and the documents at hand, they can only identify *hypotheses about potential problems*. For instance, a certain thesaurus term is very rarely used to annotate documents from the corpus. This can be due to a problem in the thesaurus (in this case it can be caused by missing synonyms) or it can just be a consequence of the fact that this particular topic is not mentioned in the corpus. The decision whether a certain observation actually reveals a problem in the thesaurus requires a human in the loop.

In this paper, we present an approach for interactively assessing the suitability of a thesaurus for automatically annotating a given document set. Our method combines the application of statistics mostly related to the notion of information contents of terms in the thesaurus and a visualization of the results of the statistical analysis in a way that helps the user to identify and further investigate potential problems in a thesaurus. We describe the method and evaluate it in two experiments with different thesauri and document sets. In particular, we take the following steps:

- We analyze two different thesauri and document sets using a combination of information theoretic measures and advanced visualization and identify some problems and non-problems
- We show that fixing the problems identified using our method actually improves automatically created annotations compared to manually created ones.

The paper is structured as follows. In section 2, we describe the methods for assessing thesauri and visualizing the result in terms of a treemap. The application of the method to the different thesauri is described in section 3. We give examples of interesting observations faciliated by the method and highlight critical problems. Section 4 describes two experiments in automatically indexing documents in the areas of medicine and economics with corresponding thesauri and compares the results to available manual annotations. We further use this experimental setup to compare the results before and after some problems have been fixed. We conclude with a discussion of open questions and future research.

2. THESAURUS ANALYSIS AND VISUALIZATION

As a first step towards supporting thesaurus analysis, we used the Collexis Engine to automatically index our document sets with the thesauri. The Collexis Engine is a state of the art system for concept-based document indexing and retrieval provided by the German company SyynX[1]. For our experiments, we used various document sets from two sources. On the one hand, we had different sets of Medline abstracts[2], ranging from around 800 randomly selected abstracts to sets of 100.000 abstracts. These abstracts were annotated with the MeSH thesaurus. On the other hand, we had a document base of about 1000 abstracts, provided by Elsevier B.V.[3], indexed with the German-English Standard Thesaurus Wirtschaft (STW).

Title Do cigarette producers price-discriminate by state? An empirical analysis of local cigarette pricing and taxation.

Authors Theodore E. Keeler, Teh-wei Hu, Paul G. Barnett, Willard G. Manning, Hai-Yen Sung

Abstract This study analyzes the interactive effects of oligopoly pricing, state taxation, and anti-smoking regulations on retail cigarette prices by state, using panel data for the 50 US states between 1960 and 1990. The results indicate that cigarette producers do price-discriminate by state, though the effect is not large relative to the final retail price. There are two further results: (1) state taxes are more than passed on - a 1-cent state tax increase results in a price increase of 1.11 cents, and (2) sellers offset state and local anti-smoking laws with lower prices, thereby blunting effects of the regulations.

Journal Journal of Health Economics

Figure 1: Example of a document abstract used for annotation

The randomly selected article shown in figure 1 is used as an example. Table 1 (a) shows the concepts found by the indexer, the manually selected keywords are shown by Table 1 (b).

(a) Automatic	(b) Manual
Concept	Keyword
Price	Pricing behavior of firms
Cigarette	Oligopoly
Panel	Effects of taxation
Regulation	Tobacco tax
State tax	Cigarette industry
Tax increase	
Retail price	
State	
Oligopoly	

Table 1: Automatically assigned vs. manually selected concepts

[1]http://www.syynx.de
[2]http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=pubmed
[3]www.elsevier.com

As we can see, there are significant differences between the index terms chosen by the human annotator and the terms selected automatically by the Collexis system. This observation is supported by the results of the evaluation discussed later. The basic question we are addressing now is whether the structure is responsible for the mismatch and whether modifications of the thesaurus can reduce the differences to the manual annotation.

2.1 Thesauri

In the experiments, we used two thesauri from different domains and with quite different characteristics in order to be able to generalize our observations. The first thesaurus is MeSH, a well established thesaurus from the medical domain that is extensively used to annotate large collections of medical documents. The second thesaurus is the German standard thesaurus for business and economics that has been created recently to provide the basis for indexing literature in the area of business and economics. In contrast to Mesh the use of this thesaurus is limited to a number of specialized libraries and there is no experience with its use for automatic indexing. In the following, we describe the two thesauri in more detail to provide the necessary background information for the analysis in section 4.

2.1.1 MeSH

The Medical Subject Headings (MeSH) thesaurus[4] is produced by the National Library of Medicine (NLM) and continuously updated since 1960. It is used for cataloging the various documents and related media and as an index to search these documents in a database and is part of the metathesaurus of the Unified Medical Language System (UMLS). This thesaurus originates from keyword lists of the Index Medicus, a comprehensive directory of medical documents, nowadays known as Medline. Medline still uses the MeSH headings as descriptors for the documents. The thesaurus, as well as the Medline database are available online. The MeSH thesaurus is a polyhierarchic thesaurus. On top, there are 16 categories, where each of is further divided into subcategories. A subcategory contains a hierarchic subtree from most general to most specific concepts (descriptors) in up to eleven hierarchical levels. These subtrees are not an exhaustive classification of the subject matter, but contain only those terms that have been selected for inclusion, representing a compromise among the needs of various disciplines and users.

Each MeSH descriptor appears in at least one place in the trees, and may appear in as many additional places as may be appropriate. In our experiments, we used the MeSH 2006 thesaurus with 31956 concepts and about 170.000 terms. The language of the MeSH thesaurus is English.

[4] http://www.nlm.nih.gov/mesh/

2.1.2 STW

The Standard Thesaurus Wirtschaft (STW)[5] is produced by GBI-Genios GmbH and the Deutsche Zentralbibliothek für Wirtschaftswissenschaften (ZBW). The thesaurus is focused on economic sciences and consists of several parts including Common Terms, Economic Science, Geography, Related Disciplines, Products, Political Economics and Industries. There is another part, primarily intended as additional access and for cross reference: the NACE concordance (Nomenclature général des activités économiques dans les Communautés Européens).

Each main part is further divided into hierarchic subparts. The concepts are polyhierarchically ordered and can be assigned to several of the subparts. So every subpart can be seen as subthesaurus. To reflect the different subthesauri, we built a single thesaurus and used the hierarchy of subthesauri as additional concepts on top of the STW concepts. The parents of a concept node are the union of broader terms and allocations. The resulting thesaurus consists of 6292 concepts and 27204 terms. The language of the STW thesaurus is German. Every concept contains one english term as synonym. The hierarchy of the subparts is only available in German.

2.2 Thesaurus Analysis

As motivated in the introduction, we believe that the quality of automatic indexing can be improved by a critical review of the thesaurus. This review consists of two basic steps: the detection of parts of the thesaurus that show an unexpected behavior and a detailed inspection of such parts resulting in a decision whether this particular part has to be revised to better support the indexing process. Literature on thesaurus creation and maintenance mentions a number of revisions that might be necessary including the following [1]:

- adaptation of the thesaurus to changes in the vocabulary of the domain of interest by means of adding of new terms
- Deletion and/or merging of rarely used terms
- splitting, extension or restriction of extensively used terms
- review of the thesaurus structure to avoid extensive subclassing

We add a new aspect to these traditional ones that arises with automatically indexing:

- identification of problematic concepts for the automatic indexer, i.e. concepts that are erroneously assigned due to misleading occurrences in the documents with improper sense

[5] http://www.gbi.de/thesaurus/

Whereas the first aspect cannot be achieved by our analysis, it supports all other revisions. In order to enable a domain expert to carry out these actions, we first have to analyze the thesaurus and detect unbalanced hierarchy structures as well as terms that are more often or less often used in indexing than we would expect. We support this step using a statistical measure that is discussed in this section. Computing such measures has also been done by other researchers in the area of thesaurus maintenance. Beside this, we also implemented a thesaurus evaluation tools that supports the identification of problematic concepts using a visualization that makes it easy for the user to spot potential problems. In the following, we introduce the measure implemented in our tool as well as the visualization techniques used.

2.3 Evaluating Thesaurus Suitability

The decision whether a term is used more often indexing as expected depends on our expectation about the frequency it should occur. This in turn depends on the level in the thesaurus hierarchy it is situated at. Generally, more common concepts should have a higher frequency than special concepts. To take this into account, we do not operate on the frequency directly, instead we propose the difference of Information Contents as a distance measure:

$$D_{IC}(c) = IC(c) - IIC(c) \tag{1}$$

where $IC(c) = -\log P(c)$ is the information content of a concept c with respect to the indexed document set, as proposed by Resnik [8].

The idea of an Intrinsic Information Content (IIC), i.e. an information content that is determined only by means of the thesaurus structure itself, was introduced by Seco et al. [9].

The IIC of a concept c denoted as IIC(c) is defined as

$$IIC(c) = -\log\left(\frac{hypo(c)+1}{max}\right) \tag{2}$$

with $hypo(c)$ as the number of hyponyms (i.e. child nodes) of a given concept c and max as the number of concepts in the whole thesaurus.

We use the Intrinsic Information Content as reference and compare it to the Information Content based on the concept frequency in the document base.

Equation 1 is related to the Kullback-Leibler Divergence used in information theory. It is defined as

$$D_{KL}(p||q) = \sum_i p(i)\log\frac{p(i)}{q(i)} \tag{3}$$

and is a measure of the differences between two probability distributions p and q. The Kullback-Leibler Divergence can be used to get an overall measure of the thesaurus suitability, instead of evaluating a single concept. It is used in this way to evaluate ontologies, examples can be found at [3, 2].

2.4 Visualizing Thesaurus Suitability

A major challenge in supporting thesaurus maintenance is to provide adequate tool support that guide the user to potential problems in a thesaurus based on the measures described above. In particular we have to find a way to provide the user with a view on the thesaurus that encodes the overall structure of (parts of) the thesaurus and the suitability values for the different terms in the thesaurus. Ben Shneiderman invented the Treemap algorithm in the early 1990s, published in [10]. With a treemap, two additional aspects can be displayed beside the thesaurus structure. One is represented by the size of the partitions, the other by its color.

Figure 2 shows the treemap of the MeSH thesaurus, where each area represents a concept in the thesaurus.

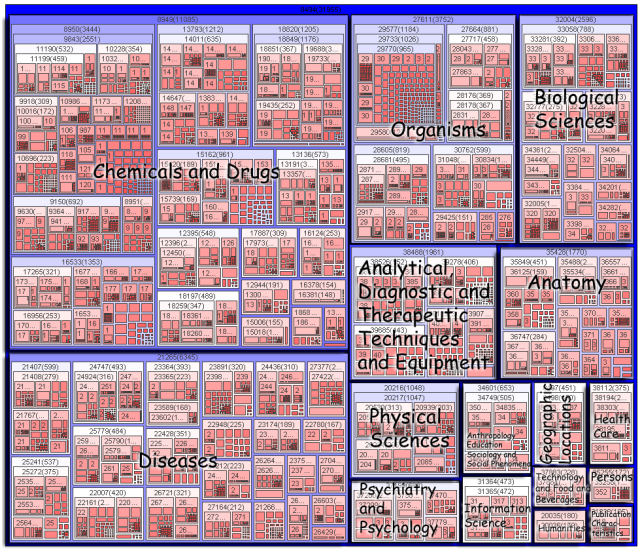


Figure 2: Treemap of the MeSH thesaurus

The hierarchy is visualized through the nesting of areas. The color of the different areas is used to represent the result of the different measures introduced above. In figure Figure 2 the color corresponds to the intrinsic information contents of a concept. While the color of concepts that are low in the hierarchy and therefore have a intrinsic high information content lean towards a red color whereas concepts with a low intrinsic information content lead towards blue. The use of this representation and in particular the color coding for representing different measures is discussed in the following section.

For our experiments, we combine combine a hierarchical common treeview with the treemap visualization. This allows interactive navigation through the thesaurus hierarchy without losing the orientation. The full interactive capabilities of the program are demonstrated in an online animated presentation at http://www.kaiec.org.

3. EXPERIMENTAL EVALUATION

We implemented a thesaurus analysis tool that supports the application of the statistical measures and the visualization introduced above. Using this tool, we analyzed the suitability of the Mesh and the STW thesaurus providing a basis for automatically indexing the document sets described above. In the following, we present the results of this analysis focussing on the evaluation of thesaurus suitability. We show that the visualization of the deviation between the intrinsic and the real information content naturally leads us to parts of a thesaurus that causes trouble in the indexing process. In particular, we discuss a number of severe indexing errors that we detected using our tool. We also show that we can improve the indexing result by locally fixing the problems detected. The evaluation is by no means exhaustive. Our goal is to show that the methods proposed in this paper works in principle. A more detailed analysis is work in progress and beyond the scope of this paper.

3.1 STW Analysis

Our analysis is based on the difference in the information content. In particular, we looked for terms with a large difference in the sense that their actual information content is lower than the expected one as this also can be caused by wrong annotations.

The IC Difference Analysis provides a very diversified picture of the thesaurus (Figure 3). The colors are calculated by the difference of the Information Content and the Intrinsic Information Content. The Intrinsic is seen as the expected information content, the IC is the actual Information Content. Red denotes concepts with a higher IC than expected, blue vice versa. Remember, that a high IC is a result of a low frequency of the concept in the document base.

The IC Diff analysis provides us with information about which parts of the thesaurus are more often used to annotate documents than one would expect. Figure 3 shows that this is the case for the subpart of the thesaurus rooted at the term *Business Economics*, which is not very surprising. In this part of the thesaurus, there are also a lot of common concepts, like *Ratio, Capital, Price, Market*. Other areas that are well represented in the document base are health care as well as finance and banking. This focus can easily be explained by looking at the journals the articles were taken from (e.g. the journal of health care economics).

In the case of *Business Economics* the high difference between expected and real information contents is not an indicator for a problem in the thesaurus but merely a result of the topics covered in the document base.

There are also cases where the difference analysis identifies problems. A good example is the term *products of paper* shown in figure 4. This term originally refers to the branch of economics concerned with the production of paper. In the indexing process, the term paper was often assigned to documents that are not concerned with this intended meaning because of phrases that contain references to scientific publications (”in a recent paper ... suggest”).

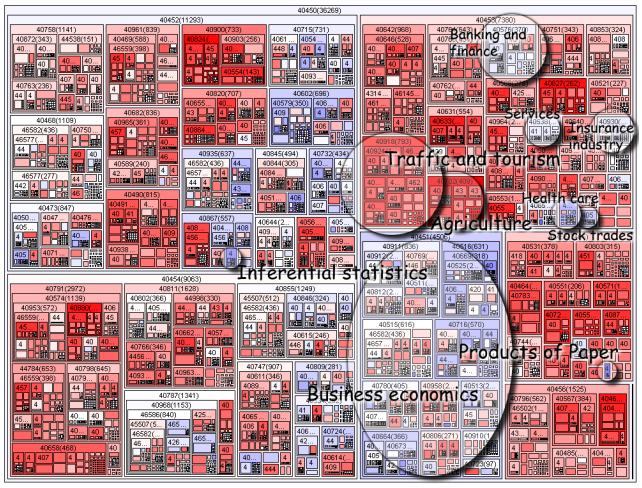


Figure 3: STW IC Difference Analydsis

The solution is to add additional information to the concept to help the indexer with the disambiguation between the two meanings of paper.

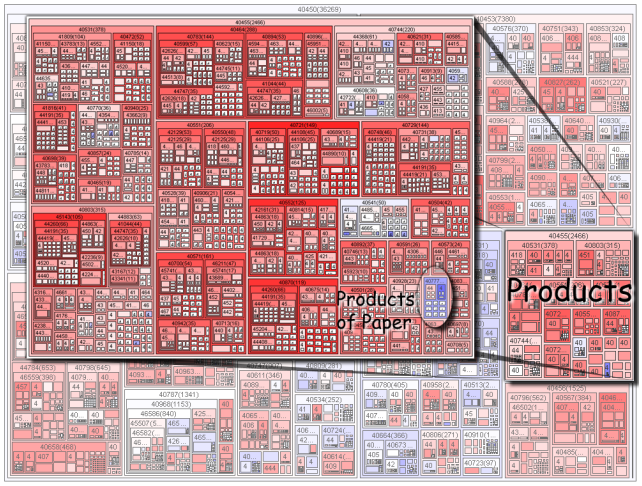


Figure 4: Zoom on Products

On the other hand, IC Diff analysis also shows parts of the thesaurus that are less often used to annotate documents that one would expect. An example would be the concept of *raw materials*. A closer look with the interactive tool reveals that the commonly used synonym *commodity* is missing in the thesaurus node. In our document base, the term *commodity* is used twice as often as *raw materials*. The terms never co-occur.

3.2 MeSH Analysis

In order to show that this kind of analysis is not only useful for newly developed thesauri, but also proves useful for analyzing well established thesauri with respect to their abil-

ity to describe a given document base, we also analyzed the MeSH thesaurus on the basis of the above mentioned Pubmed abstracts.

A first interesting results from the analysis of the MeSH Thesaurus is concerned with the structure of the MeSH thesaurus. Figure 5 shows the Difference of Information content for the whole thesaurus. The MeSH Category *Organisms* is highlighted.

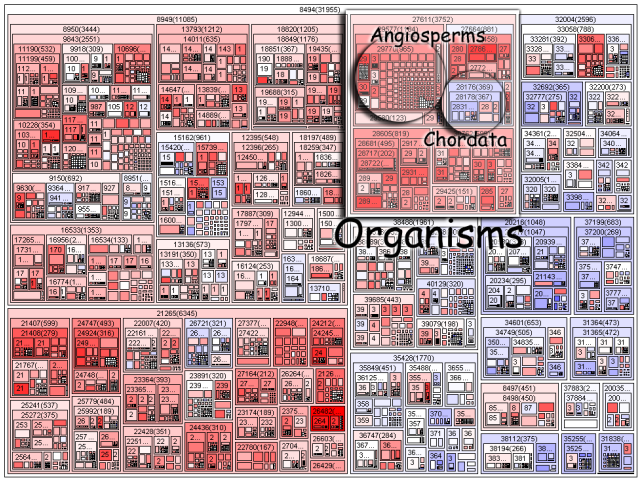
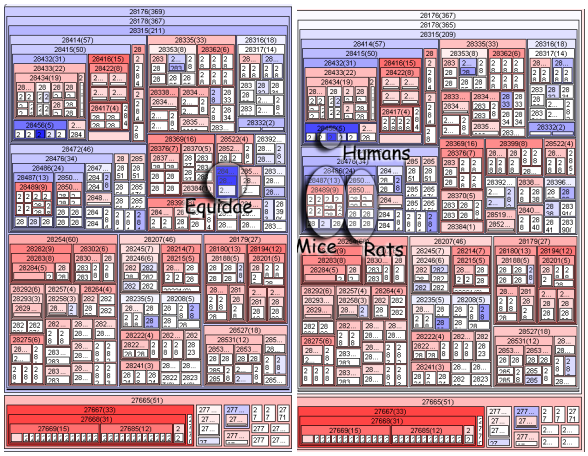


Figure 5: MeSH IC Difference Analysis

First of all, you see the *Angiosperms* concept with lots of subconcepts. The structure is visibly different to the other concepts. Whereas such a structure could indicate a problem with the thesaurus, in this case, it just reflects the fact that "the angiosperms, or flowering plants, are one of the major groups of extant seed plants and arguably the most diverse major extant plant group on the planet, with at least 260,000 living species classified in 453 families" [6] This clearly shows that the structural irregularity in the thesaurus is not a design mistake but correctly reflects the nature of the domain.

When looking further at the differences in the information content another interesting concept sticks out: The Chordata (a group of animals including the vertebrates and some closely related invertebrates). It shows a significant lower Information Content as a result of some concepts with a unexpected high frequency compared to the other concepts in this group. Figure 6(a) zooms into the *Chordata* concept for further analysis. We see several blue areas and two dark blue concepts. One of them are the *Equidae*, also know as horse-like animals. The result shows a by far too low information content for this concept, thus it has a very high frequency in the document base. As the document base is not dedicated to horse diseases, this indicates a problem: In the MeSH Thesaurus, we find the following terms for the concept *Equidae*: Asses, Donkeys, Mules, Zebras, Ass, Equus asinus (the horses themselves are contained in a subconcept of Equidae). The problem lies in the term *Ass*, as the language normalizer interpreted every occurrence of the word *as* as the singular form of *Ass*. This example nicely illustrates the problems that can arise in the course of automatic indexing and it also shows that our method is an adequate means to identify such potential problems.



(a) With Equidae (b) Without Equidae

Figure 6: MeSH IC Difference Analysis Animals

We fixed this issue and rerun our analysis, the result is shown by Figure 6(b). The *Equidae* are now as rare as they should be. But there are still several blue concepts. The most frequent concept now is the concept *Humans*, which is not very surprising as most articles are concerned with the treatment of human patients. Other high frequent concepts are *Mice* and *Rats*, which gives us an direct insight on the favorite subjects of animal testing for drug discovery.

We conclude that these derivations of the information content do not point to a problem in the thesaurus or the indexing process as they are a result of the nature of the domain of interest.

4. EVALUATION OF THE AUTOMATIC INDEXING

The general idea for the evaluation is a comparison of the automatically attributed concepts with a gold standard of keywords that have been manually selected by human experts. We expect, that the indexing process finds at least the same keywords as selected by humans, if the keyword appears in the abstract. Additional concepts should be found due to the fact, that every concept appearing in the abstract is found.

For this evaluation, we used two document corpora. On the one hand, there was a randomly selected collection of 706 Medline abstracts [7]. These articles were annotated with the MeSH thesaurus.

[6] http://tolweb.org/Angiosperms/20646/2005.06.03 in The Tree of Life Web Project, http://tolweb.org/

[7] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=pubmed

On the other hand, we had a document base of 391 economic abstracts, provided by Elsevier B.V.[8], indexed with the German-English Standard Thesaurus Wirtschaft (STW). We fetched manually selected STW keywords from the Econis Database of the German central economic library [9].

4.1 Generalized Precision and Recall

It has been widely acknowledged that traditional definitions of precision and recall have serious limitations with respect to measuring the overlap between terms from structured vocabularies like thesauri. Different researchers have proposed alternative measures, often referred to as generalized precision and recall that do not only take the overlap between concepts into account but also consider the semantic distance between between concepts that are not in the common term set [5, 7]. The most recent proposal for generalizing precision and recall has been made by Euzenat [4], who defines generalized precision and recall in the following way:

$$Prec_\omega(A,R) = \frac{\omega(A,R)}{|A|} \qquad Rec_\omega(A,R) = \frac{\omega(A,R)}{|R|} \quad (4)$$

Here A is the automatically created annotation, R is a reference annotation, in our case the manually created one and ω is a function that measures the overlap between A and R.

There are many options for choosing ω. In the context of comparing annotations from a thesaurus it makes sense to base the definition on notions of semantic similarity between concepts. There are a number of proposals for semantic similarity measures including purely structural measures, as well as measures that are based on information theoretical concepts. A measure that tries to combine the two is proposed in [9] by means of the above mentioned Intrinsic Information Content. We performed a comparison of these different measures and evaluated their ability to adequately capture the quality of an annotation. Based on this evaluation, we selected the measure of Lin with Intrinsic Information Content as a basis for our experiments [6]. We decided in favour of the Intrinsic Information Content due to the relatively small document sets. With larger document corpora, probably the traditional Information Content should be used.

According to this measure, the similarity of two terms is computed as follows:

$$Similarity_L(c_1,c_2) = \frac{2 \cdot IIC(LCS)}{IIC(c_1) + IIC(c_2)} \quad (5)$$

LCS denotes the first common term reached when moving from c_1 and c_2 upwards in the hierarchy.

[8] www.elsevier.com

[9] Deutsche Zentralbibliothek für Wirtschaftswissenschaften, http://zbwopc4.zbw.ifw-kiel.de:8080/DB=1/LNG=DU/

This leads us to the following definition of precision and recall with respect to a single document:

$$Recall = \frac{\sum_{r \in R_d} \max_{a \in A_d} Similarity_L(r,a)}{|A_d|} \quad (6)$$

$$Precision = \frac{\sum_{a \in A_d} \max_{r \in R_d} Similarity_L(a,r)}{|R_d|}. \quad (7)$$

where A_d refers to the set of automatically created annotations for document d, R_d and to the set of manually assigned keywords of document d. Note that we can calculate the overall precision and recall for a complete document set by summing over all documents d.

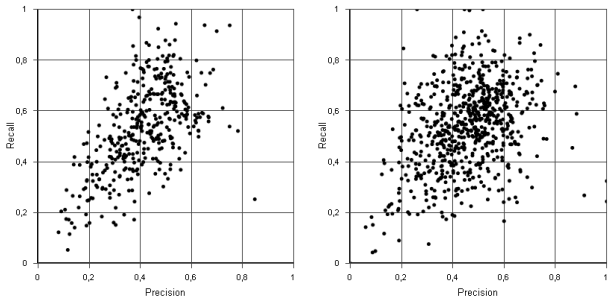
4.2 Results

Table 2 summarizes the result of the automatic annotation experiments. The numbers given correspond to the overall precision and recall over the complete document set.

Document Base	# Keywords	∑Similarity	Recall
STW/Elsevier	1646	862	0.52
MeSH/Medline	8143	4197	0.52
	# Concepts	∑Similarity	Precision
STW/Elsevier	3377	1331	0.39
MeSH/Medline	10041	4353	0.43

Table 2: Generalized Precision and Recall

Figure 7 provides a more detailed view of the results in terms of the precision and recall of individual documents in the set.



(a) STW Lin (b) MeSH Lin

Figure 7: Generalized Precision and Recall

Looking at the results, we see that the overall results are more or less comparable for the Mesh Thesaurus that can be assumed to have a high quality and the STW thesaurus where we do not have any information about the quality. In total, we can also note that there is a lot of space for improvement in both cases in particular with respect to precision. This observation is supported by the detailed results shown in figure 7. In particular for the STW experiment we can see that almost none of the documents reach a precision of more than 70 percent.

4.3 Improving the Automatic Indexing Result

As stated in the introduction, the ultimate goal of the thesaurus analysis is an improvement of the indexing results in terms of increased precision and recall. The development of a comprehensive method for repairing indexing results is a topic of ongoing research. Nevertheless, we performed some initial experiments in improving the indexing results based on the results of the thesaurus analysis. In particular, we improved concepts that have been identified as problematic on the basis of a high difference between expected and real information content from the index terms and re-calculated precision and recall. In the case of the STW thesaurus, we removed the ambiguous synonym *paper* from the concept. This lead to an increase of the precision by 4.8%.

In the case of the MeSH example, improving the normalizer led to an increase of the precision by 1.2% while the recall remained unchanged. These improvements are not too impressive at first sight, but we have to keep in mind that the increase is the result of just fixing a single problem of one out of about 32000 concepts.

5. DISCUSSION AND CONCLUSIONS

In this paper we present a method for analyzing the suitability of a thesaurus with respect to providing the basis for automatically indexing a given document set. We illustrated the methods based on results of an automatic indexing experiments including thesauri and documents from two different domains. We showed that based on a suitable visualization of some statistical measures for the suitability of a thesaurus it is easy to identify potentially problematic parts of a thesaurus and that a manual inspection of these problematic parts often reveals problems that appeared in the indexing process. We also presented results of initial experiments in improving annotations based on the result of the analysis in which fixing a single problem leads to a significantly increased precision of the annotation.

We conclude that interactive thesaurus assessment is a suitable mean to improve the results of automatic document annotation. Especially for relatively small document sets identifying indexing errors can have a significant impact on the quality of the annotation. Apart from this general observation, there are a number of issues that need further investigation. One open question is the impact the choice of a particular distance measure as the basis for generalized precision and recall has on the results of our analysis. We believe that a comprehensive theory of generalized precision and recall is needed to answer that question. Another important observation was that many of the problems found by our method can actually be seen as problems of the indexing algorithm rather than problems of the thesaurus. This means that the method can also be used to analyze indexing methods.

There are several directions for future work that are implied by this research. The major direction of research are methods for improving annotation results based on the results of the analysis. So far, we have only considered the improvement of over-represented concepts in order to improve the precision of annotations. In order to improve recall, we need to identify terms that are under-represented in the annotations and find out why this is the case. A possible problem is the lack of proper synonyms. Finding such missing synonyms is a possible way to improve recall that will be investigated in the future.

Please refer to http://www.kaiec.org for high resolution media of all figures used in this paper, as well as an animated presentation of the interactive tool.

6. REFERENCES

[1] M. Burkart. *Grundlagen der praktischen Dokumentation und Information, Band 1*, chapter Thesaurus, pages 151–153. Saur, 2004.

[2] J. Calmet and A. Daemi. Assessing conflicts in ontologies. Technical report, IAKS Calmet, University Karlsruhe (TH), Germany, 2004.

[3] J. Calmet and A. Daemi. From entropy to ontology. Technical report, Institute for Algorithms and Cognitive Systems (IAKS), University of Karlsruhe (TH), Germany, 2004.

[4] J. Euzenat. Semantic precision and recall for ontology alignment evaluation. In *IJCAI*, pages 348–353, 2007.

[5] U. Hahn and K. Schnattinger. Towards text knowledge engineering. In *AAAI/IAAI*, pages 524–531, 1998.

[6] D. Lin. An information-theoretic definition of similarity. In *Proc. 15th International Conf. on Machine Learning*, pages 296–304. Morgan Kaufmann, San Francisco, CA, 1998.

[7] D. Maynard, W. Peters, and Y. Li. Metrics for evaluation of ontology-based information extraction. In *WWW 2006 Workshop on "Evaluation of Ontologies for the Web" (EON)*, Edinburgh, Scotland, 2006.

[8] P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI-95)*, 1995.

[9] N. Seco, T. Veale, and J. Hayes. An intrinsic information content metric for semantic similarity in wordnet. In *Proceedings of the 16th European Conference on Artificial Intelligence*, pages 1089–1090. Valencia, Spain, 2004.

[10] B. Shneiderman. Tree visualization with tree-maps: 2-d space-filling approach. *ACM Trans. Graph.*, 11(1):92–99, 1992.