Assessing Thesaurus-Based Annotations for Semantic Search Applications

Kai Eckert*, Magnus Pfeffer and Heiner Stuckenschmidt

Computer Science Institute
University of Mannheim
A5, 6 68159 Mannheim, Germany
Email: kai@informatik.uni-mannheim.de
Email: pfeffer@bib.uni-mannheim.de
Email: heiner@informatik.uni-mannheim.de
*Corresponding author

Abstract: The use of thesaurus-based indexing is a common approach for improving the result of document retrieval. With the growing amount of documents available, manual indexing is no longer a feasible option and statistical methods for automated document indexing are becoming 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 and as a basis for the specific indexing method used is of crucial importance in automatic indexing. The reason being that there is no human in the loop that can spot and avoid indexing errors. We propose the use of an interactive tool 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 the evaluation of indexing results. The examples show that the tool supports the detection and correction of errors, leading to a better indexing result.[1]

Keywords: Content-Based Retrieval, Thesaurus Quality, Evaluation, Information Content, Visualization

Biographical Notes: Kai Eckert received his masters degree in computer science (German equivalent: Diplom-Informatiker) from the University of Mannheim in 2004. Currently he is working as PhD student with main focus on semantic search applications and the evaluation of automatic document annotations.

Magnus Pfeffer received his masters degree in computer science (German equivalent: Diplom-Informatiker) from the University of Kaiserslautern in 2003. He received an additional masters degree in library and information science from the Humboldt-University of Berlin in 2007. Currently he is working at the Mannheim University Library as the vice head of the IT department. His research focus is on automatic classification and thesaurus-based search.

Heiner Stuckenschmidt is professor for knowledge representation and knowledge management at the University of Mannheim, Germany. His research group is working in the area of knowledge representation in distributed and heterogeneous environments such as the semantic web as well as on the application of semantic web technologies in scientific information systems. Before moving to Mannheim, he was post-doc researcher in the Knowledge Representation and

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

Reasoning Group at the Vrije University Amsterdam where he received his PhD for a thesis on Ontology-based Information Sharing. He is author of a book on the same topic and editor of two other books on related topics.

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 information retrieval methods.

In particular, they solve the problem of synonyms by explicitly representing information about synonymous terms and relating them to a preferred term that is used to describe the content of a document. Thesauri also allow the disambiguation of homonyms by expressing the homonymous concepts through different and expanded thesaurus entries.

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 continually increasing amount of documents, manual annotation in a timely fashion is not feasible any more. Therefore, techniques for automatically annotating documents with terms from a thesaurus have received significant attention recently (cf. Clifford Gay and Aronson (2005); Stuckenschmidt et al. (2004); Ferber (1997)). Figure 1 presents the architecture of a typical thesaurus-based semantic search system. Such a system consists of two parts. The first one implements the actual search and retrieval functionality that takes a user query as input, compares it to document annotations and computes a ranked set of results. The second part of a semantic search system that is in the focus of our work is concerned with the creation of document annotations. This process relies on three different components

- the document set to be annotated
- the thesaurus that provides index terms
- the indexing method used and its ability to deal with ambiguities in the text.

With the increased use of automatic indexing techniques arises a need to assess the quality of their results. We argue that this need is best met by tools that enable a human expert to analyze all of the components of the annotation system mentioned above. While the documents and the thesaurus can directly be assessed, the indexing method is best evaluated based on its results in regard to the index terms chosen. With this approach, we want to keep the human expert in the loop of an automatic annotation system and use the expert knowledge in the most efficient way to ensure the quality of the whole information retrieval system.

In this article, we present an approach for interactively assessing thesaurus-based annotation systems. Our method combines the application of statistics mostly related to the notion of information content of terms in the thesaurus and a visualization of the results

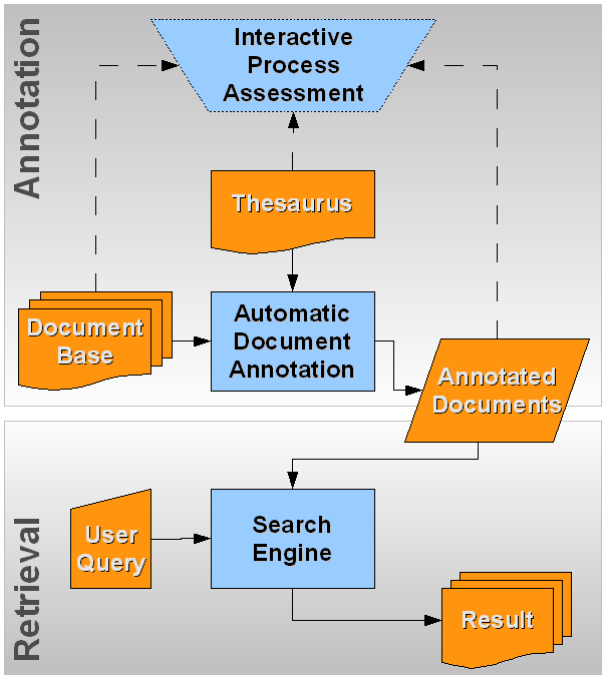


Figure 1 Schematic view of a supervised annotation and retrieval process

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 experiments with a well-known thesaurus and document set. In particular, we take the following steps:

- We present a real world annotation problem from the medical area as a basis for our investigation (section 2).
- We describe the information theoretic measures that we developed for thesaurus assessment and show the treemap visualization that is a central aspect of our analysis approach (Section 3).
- We present the application of our approach to the annotation problem introduced in section 2, give examples of interesting observations facilitated by the method and highlight critical problems (Section 4).
- We present an evaluation approach based on the Precision and Recall of the annotating system with respect to the manually selected keywords and show that fixing the problems identified using our method actually improves the annotation results (Section 5).
- We conclude with a discussion of open questions and future research. (Section 6).

2 An Example Problem

Before we discuss details of our method for assessing indexing results, we first introduce an example indexing problem in order to clarify the challenge addressed in our work.

2.1 Data and Indexing Method

We chose an annotation problem from the medical domain as a basis for our investigations. In particular, we look at the task of indexing abstracts from the Medline Digital Library[a] with terms from the MeSH thesaurus[b].

For our experiments, we used a document set containing 7376 randomly selected Medline abstracts. Only abstracts that are already annotated with keywords from the MeSH thesaurus by domain experts were used. These manually assigned keywords provide us with a reference for assessing the quality of automatic index results.

We indexed these documents automatically using the *Collexis Engine*, a state-of-the-art integrated system for concept-based document indexing and retrieval provided by the German company SyynX[c]. The engine has already been applied successfully in the medical area and therefore provides an adequate basis for our investigations.

Title Analysis of the main factors affecting the evaluation of the radon dose in workplaces: the case of tourist caves.

Authors Carlos Sainz, Luis Santiago Quinds, Ismael Fuente, Jorge Nicols and Luis Quinds

Abstract High concentrations of radon exist in several workplaces like tourist caves mainly because of the low ventilation rates existing at these enclosures. In this sense, in its 1990 publication, the ICRP recommended that high exposures of radon in workplaces should be considered as occupational exposure. In developed caves in which guides provide tours for the general public great care is needed for taking remedial actions concerning radon, because in some circumstances forced ventilation may alter the humidity inside the cave affecting some of the formations or paintings that attract tourists. Tourist guides can work about 1900 h per year, so the only option to protect them and other cave workers from radon exposure is to apply an appropriate system of radiation protection (Term: Radiation Protection, Cutoff: 0) mainly based on limitation of exposure by restricting the amount of time spent in the cave. Because of the typical environmental conditions inside the caves, the application of these protecting actions requires to know some indoor air characteristics like particle concentration, as well as radon progeny behaviour in order to get more realistic effective dose values In this work the results of the first two set of radon measurements program carried out in 10 caves located in the region of Cantabria Spain are presented.

Journal Journal of Hazardous Materials

Figure 2 Example of a document abstract used for annotation

Figure Figure 2 shows an example of a medline abstract. Table 1 (a) shows the concepts found by the automatic indexer, the manually selected keywords are shown by Table 1 (b).

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.

To avoid ambiguity, in this article we will use *keyword* for index terms chosen by human annotators and *concept* for the ones found by the automatic indexing system.

2.2 The MeSH Thesaurus

In the experiments, we used the MeSH thesaurus, a well established thesaurus from the medical domain that is extensively used to annotate large collections of medical documents. In the following, we describe the thesaurus in more detail to provide the necessary background information for the analysis in Section 4.

[a] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=pubmed
[b] http://www.nlm.nih.gov/mesh/
[c] http://www.syynx.de

(a) Automatic

Concepts (Rank)
Radon (1)
Ventilation (0,33)
Work (0,33)
Workplace (0,33)
Affect (0,16)
Air (0,16)
Health Services Needs and Demand (0,16)
Humidity (0,16)
Radiation Protection (0,16)
Spain (0,16)
Occupational Exposure (0,16)

(b) Manual

Keywords
Air Pollutants, Occupational
Air Pollutants, Radioactive
Environmental Exposure
Humans
Leisure Activities
Occupational Exposure
Radon
Spain

Table 1 Automatically assigned vs. manually selected concepts

The Medical Subject Headings (MeSH) thesaurus[d] 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.

It consists of the three parts

1. MeSH Tree Structures,

2. MeSH Annotated Alphabetic List and

3. Permuted MeSH.

The Tree Structures contains various subtrees of descriptors. 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.

For each appearance of a descriptor, a number is assigned, like in Figure 3. These numbers are used to locate the descriptors in each tree and to alphabetize those at a given tree level and have no intrinsic significance; e.g., the fact that D12.776.641 and D12.644.641

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

MeSH Heading	Ethics
Tree Numbers	F01.829.500.519 K01.316 K01.752.256 N05.350
Scope Note	The philosophy or code pertaining to what is ideal in human character and conduct. Also, the field of study dealing with the principles of morality.
Entry Terms	Egoism Ethical Issues Metaethics Moral Policy Natural Law Situational Ethics
Allowable Qualifiers	CL HI
Unique ID	D004989

Table 2 Example of a MeSH Descriptor

both have the three digit group 641 does not imply any common characteristic. The numbers are subject to change when new descriptors are added or the hierarchical arrangement is revised to reflect vocabulary changes.

- Abnormalities C16.131
  - Abnormalities, Drug Induced C16.131.42
  - Abnormalities, Multiple C16.131.77
    - Alagille Syndrome C16.131.77.65
    - Angelman Syndrome C16.131.77.95

Figure 3 Excerpt of the MeSH structure

Table 2 shows an example of a MeSH Descriptor. The MeSH Heading is followed by several tree numbers denoting the multiple positions in the different subtrees of the MeSH thesaurus. A free scope note is used to describe the heading to the user. The different synonyms for the heading are described by the entry terms. One can use qualifiers to narrow the heading in a search application. And at last there is a unique ID for each heading.

3 Thesaurus Analysis and Visualization

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 tasks that might be necessary including the following taken from Burkart (2004):

- 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 chose to add another aspect to these traditional ones that deals with automatic indexing:

- identification of problematic concepts for the indexing software, i.e. concepts that are erroneous assigned or missing

Our analysis cannot help a human expert with the first task, but it can support the remaining four tasks.

In order to enable a domain expert to carry out these actions, we 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. 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.

3.1 Statistical Analysis

The decision whether a term is used more often for 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. 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 (referred to as IC Difference Analysis):

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

where $IC_a(c) = -\log P_a(c)$ is the information content of a concept c with respect to the automatically annotated document set, as proposed by Resnik (1995). $P_a(c)$ is determined by the frequency of documents that are annotated with concept c.

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. (2004).

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 Calmet and Daemi (2004b,a).

The IC Difference Analysis is not limited to Equation 1. In this paper, we use another two variants that depend on the existence of manually selected keywords:

First of all, Equation 1 can be applied to the manually selected keywords as well. In this case, the Information Content is based on the frequency of documents that are manually annotated with the given keyword:

$$D_{IC_m}(k) = IC_m(k) - IIC(k) \tag{4}$$

If we replace the IIC and use the manual selected keywords as reference, we get a very interesting analysis that directly shows the difference of the automatic annotations to the manually selected ones:

$$D_{IC_{am}}(c) = IC_a(c) - IC_m(k) \tag{5}$$

This measure shows deviations between manually and automatically assigned keyword and therefore directly points to potential problems in the automatic indexing process.

In principle the measures in Equation 1, 4 and 5 are used for the same purpose: to detect deviations between the actual and the expected information content of a concept. Based on the information available (thesaurus structure, manual annotations, automatic annotations) and the target of the evaluation (manual or automatic annotations) the different measures can be used for this purpose.

3.2 Interactive Visualization

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 the thesaurus or selected parts of it and the evaluation results for the different concepts in the thesaurus. Ben Shneiderman tried to get an overview of disc usage of a particular hard drive and needed a compact representation of its directory structure, showing additional information like file size and file type in one view. He invented the treemap algorithm in the early 1990s, published in Shneiderman (1992):

According to Shneiderman, treemaps are a representation designed for human visualization of complex traditional tree structures: arbitrary trees are shown with a 2-d space-filling representation. Consider a tree with weight or size information attached to each node and a 2-d space with corners (x_1, y_1) and (x_2, y_2). For each child of the root node, a partition of the space along the x-axis is calculated. For the first partition, this reads as

$$x_3 = x_1 + \left(\frac{|c_1|}{|r_1|}\right)(x_2 - x_1) \tag{6}$$

with $|c_1|$ as the size of child node 1 and $|r_1|$ as the size of the root node. For the next level, the corresponding partition is partitioned again along the y-axis, then again on the x-axis and so on. Shneiderman called this approach the "slice-and-dice" algorithm. Since then, a lot of different implementations and optimizations were made by several people. One such optimization are so-called squarified treemaps, which try to avoid the long and thin rectangles resulting from small nodes. More sophisticated approaches are presented in Shneiderman and Wattenberg (2001) and Bederson et al. (2002).

Marc Smith and Andrew Fiore used the treemap algorithm to visualize the Usenet newsgroups Smith and Fiore (2001). Their treemap component is used in our evaluation tool. 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[e].

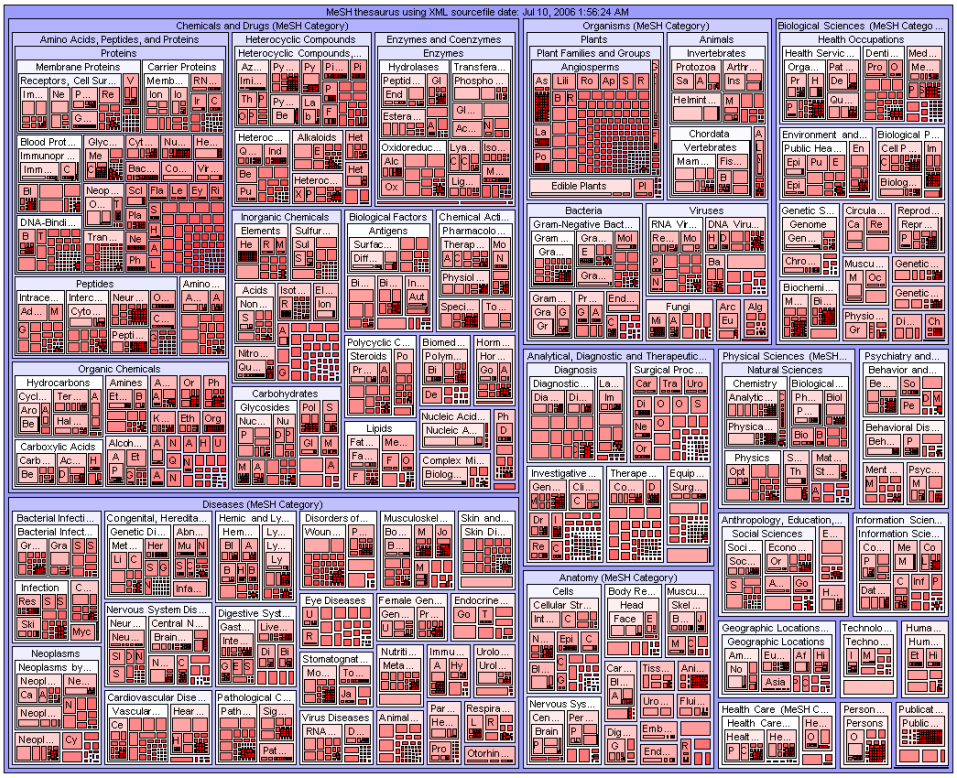


Figure 4 Treemap of the MeSH thesaurus (colours represent the IIC of concepts)

Figure 4 shows the treemap of the MeSH thesaurus, where each area represents a concept in the 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 4 the color corresponds to the intrinsic information content 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

[e] As the visualization relies on the use of colors as an integral part, the figures in the remainder of this paper should be viewed in full colors as otherwise important information is lost. Full color versions of all figures are provided at http://www.kaiec.org

lean towards blue. The use of this representation and in particular the color coding for representing different measures is discussed in the following section.

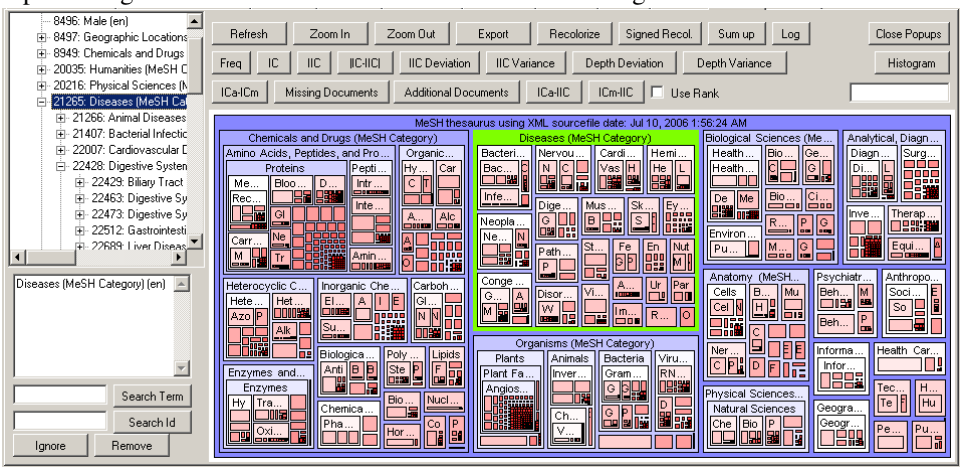


Figure 5 Interactive tool

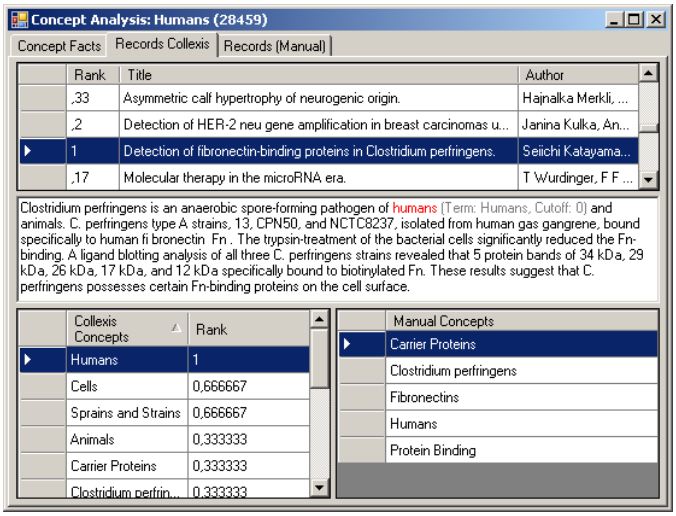


Figure 6 Concept Analysis

The treemap visualization requires some time for the user to get familiar with. A major drawback of treemaps is the possibility for the user to lose the orientation in the hierarchy as the visualization can not provide information about the environment of the currently selected top concept. To deal with this problem, our prototype (Figure 5) combines a hierarchical common treeview with the treemap visualization. This allows interactive navigation through the thesaurus hierarchy without losing the orientation.

The tool implements the difference measure mentioned above as well as a number of additional measures including the frequency of terms and some statistical measures on deviation and variance of Information Content. By double-clicking on the treemap or by selecting a concept in the treeview, the user can zoom into the thesaurus structure. For a selected concept, a lot of information is provided:

- The synonymous terms associated with this concept

- The number of child concepts
- The information content of this concept based on automatic and manual annotations
- The intrinsic information content based on its position in the thesaurus hierarchy
- Term Frequency in the whole document base
- Document Frequency based on automatic and manual annotations
- Lists of all documents that are annotated with this concept (automatic and manual)

Figure 6 shows such a list for an example concept. If a specific document is selected, the abstract of the document and lists of concepts the document is annotated with are shown. In the abstract, possible locations for term occurrences are highlighted, based on a heuristic approach.

The full interactive capabilities of the program are demonstrated in an online animated presentation at `http://www.kaiec.org`.

4 Real-World Applications

On its own, the treemap visualization provides an overview of the thesaurus structure only. When combined with the analytical measures, a powerful means to assess the result of an annotation process emerges. Its strength is that there is no need to have two sets of annotations for an analysis. The intrinsic information content of the thesaurus nodes can be used as a reference instead.

In the following subsections we will describe the use of our analysis tool to tackle three distinct tasks that are representative for assessing indexing results. We will illustrate these workflows with examples from an actual analysis performed on the document base described in Section 2.1.

The involved analyses of Section 3.1 and their correspondence to the described tasks are visualized in Figure 7. In our scenario, we can use three data sources:

- The thesaurus tree itself,
- the manually assigned annotations, if available, and
- the automatically assigned annotations.

For each concept in the thesaurus, we calculate the different information contents: IIC, IC_m and IC_a. Depending on the task, we choose either IC_m or IC_a as value of interest and compare it to a reference value, which can be either IIC or IC_m.

4.1 Analyzing Manual Annotations

Currently, thesauri are primary designed and used for manual annotation of documents. So there are many bibliographic databases or document sets that have been meticulously annotated by domain experts. Analyzing such an annotated document base with the tool presented in this article can help finding problems with the thesaurus structure and its keywords.

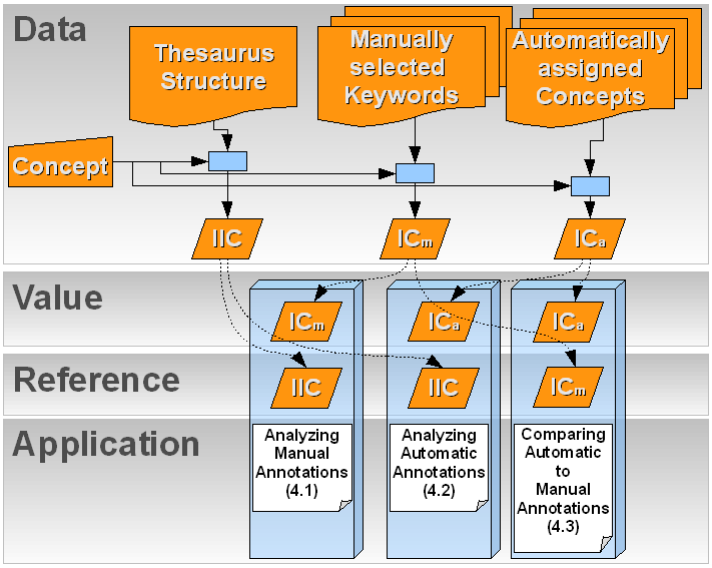


Figure 7 Schematic view of the evaluation methods

Here, the information content of the keywords is compared to their intrinsic one as described in Equation 4. The intrinsic information content implicates that the thesaurus is hierarchically structured with child nodes being more specialized and parent nodes being more general and that a keyword should occur less often in the document base, the more specialized it is.

The result of the comparison is used to colorize the respective nodes in the treemap visualization. Blue is used for positive numbers, red for negative ones while the color intensity correlates with the absolute difference. Blue tiles thus represent keywords that have been used more than their place in the thesaurus hierarchy would indicate and red tiles those that have been used less. Both cases could be a cause for concern, as the thesaurus is not only used for annotation, but to facilitate searching the document base using its keywords. Usually, the query is refined by specializing and generalizing its keywords, until a satisfactory result set is achieved. For queries containing ”blue” keywords, whose corresponding thesaurus node have no children, the result set cannot be minimized in this way. The opposite holds true for ”red” keywords whose nodes have no parents.

Figure 8 shows the result of this analysis for the example document base. On first glance, the node representing the keyword *Angiosperms* with its subconcepts is visibly different. While such a structure could indicate a problem with the thesaurus, in this case, it 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”[f] While the structural irregularity of the thesaurus correctly reflects the nature of the domain, the deep red color indicates that the individual keywords are used very infrequently.

Our next example is right next to the previous one: The node representing the keyword *Chordata* (a group of animals including the vertebrates and some closely related

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

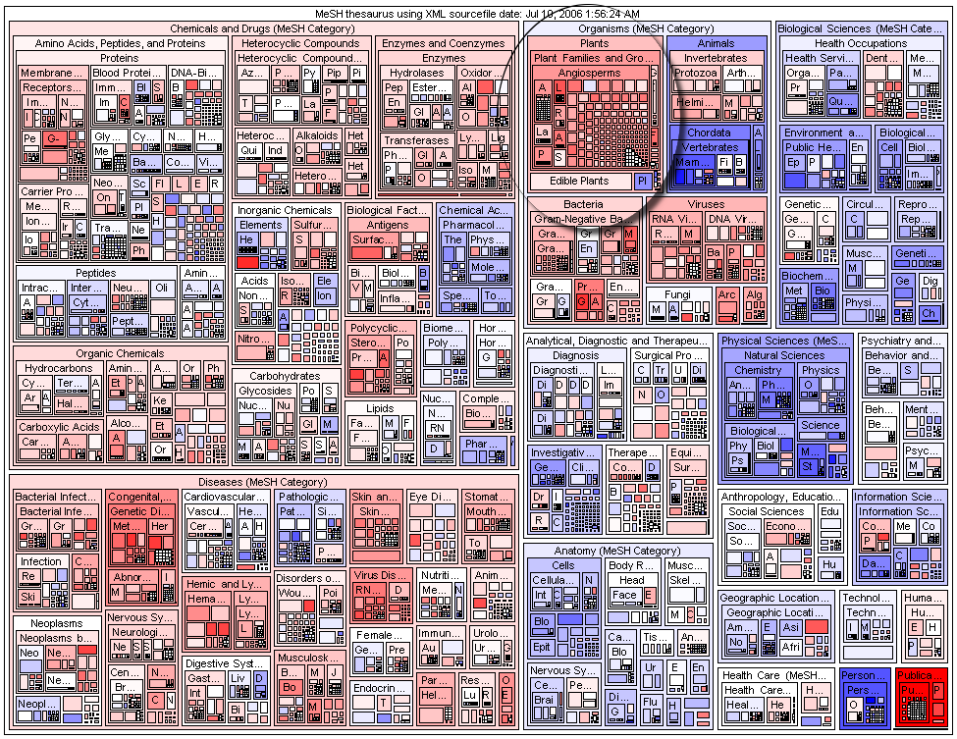


Figure 8 MeSH IC Difference Analysis ($IC_m - IIC$) on manually selected keywords

invertebrates). Its dark blue color is a result of keywords which show an unexpected high frequency in the document base.

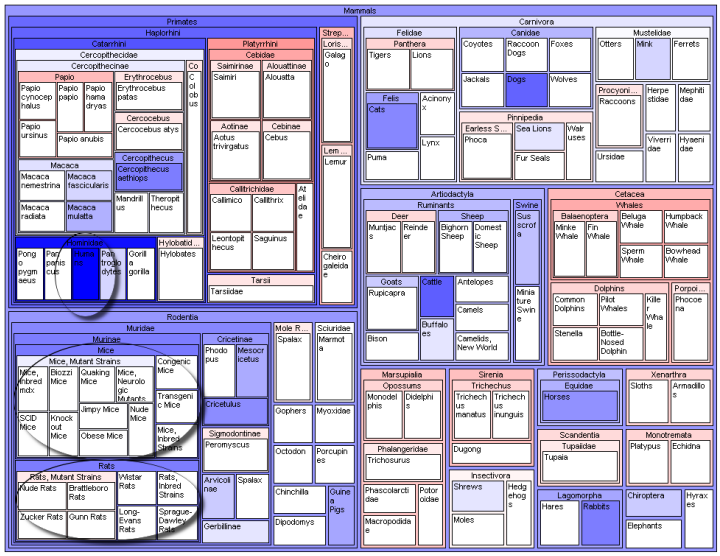


Figure 9 MeSH $IC_m - IIC$ Analysis of the node representing *Mammals*

Figure 4.1 zoomes into its child node *Mammals*. The keyword with the lowest information content is *Humans*, which is not very surprising as most Medline articles are concerned with the treatment of human patients. Others are *Mice* and *Rats*, which gives us an direct insight on the favorite subjects of animal testing for drug discovery. All these keywords are among "check tags" in the Medline database that are explicitely reviewed for every article and so have a much higher occurance in our sample document base.

The examples in this sections show that our tool is able to pinpoint problematic keywords for a given thesaurus and document base. As the developers of search applications have to take both the structural characteristics of the thesaurus and the annotation conventions of the document set into account, the $IC_m - IIC$ analysis should be of some value when designing the search interface.

4.2 Analyzing Automatic Annotations

When new document sets need to be annotated using existing thesauri, automatic processes are increasingly considered. Analyzing an automatically annotated document set with the tool presented in this article can help finding keywords that unsettle the indexer.

Here, the information content of the concepts is compared to their intrinsic one as described in Equation 1. The result is again used to colorize the treemap visualization in the way described in the previous subsection. Structural irregularities can still be spotted, and deeply tinted tiles represent possibly problematic concepts. But the root of the problem can now be either the thesaurus itself, the indexer or a combination of both. Typical problems include:

Context Dependence Concepts are sometimes homonyms of commonly used terms in a text. This preferably happens in highly specialized domains where special terminology is used. In this case there are two options. Either, advanced mechanisms for context detection can be used or in cases where these methods are too expensive, the corresponding term can be deleted from the thesaurus to avoid false annotations. This kind of problem normally causes a relatively low information content with respect to the automatic annotations.

Missing Definitions Sometimes, concepts are not detected in documents because a certain synonym used in the text is not included in the thesaurus. In this case, the definition of the concept has to be refined adding the corresponding synonyms. This problem normally causes a relatively high information content with respect to the automatic annotations.

Normalization Errors In cases where rather simple linguistic tools are used for preprocessing the meaning of terms can be lost as ambiguity is introduced in the normalization step. In this case we either have to use more advanced preprocessing methods that are capable of eliminating the ambiguity introduced for instance by first detecting noun phrases and only using them as a basis for indexing. As above, if this approach is too expensive, we can also eliminate the corresponding terms from the thesaurus to avoid wrong annotations. This kind of problem normally causes a relatively low information content with respect to the automatic annotations.

Indexing Preferences Human annotators sometimes show certain preferences in selecting index terms that cannot be reproduced in an automatic indexing process. A typical

example is the use of check-tags, predefined lists of index terms that can more easily be assigned by selecting a check-box. These terms will be over-represented in manual annotations. These terms should be treated separately in the indexing process and special strategies need to be developed for this purpose. This problem normally causes a relatively high information content with respect to manual annotations.

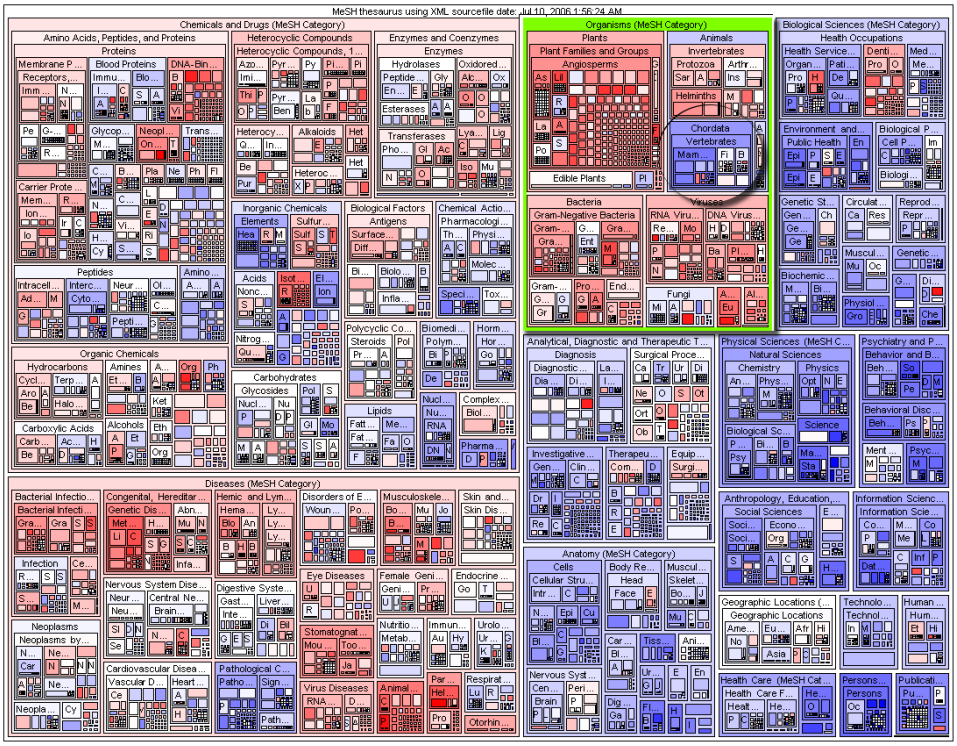


Figure 10 MeSH IC Difference Analysis ($IC_a - IIC$) on automatically selected concepts

Figure 10 shows the result of this analysis for the example document base. The tile representing the concept *Chordata* is standing out deep blue. Figure 4.2 zooms into the *Mammals* concept for further analysis. We see several blue areas and two dark blue concepts. One of them are the *Equidae*, also known 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 itself are contained in a subconcept of Equidae). The problem lies in the term *Ass*, as the language normalizer used by the indexer interpreted every occurrence of the word *as* as the singular form of *Ass*.

Another deep blue tile can be found within the tile representing the node *Biological Sciences* (Figure 12): *Attention*. In this case, the problem is not as obvious as in the case of *Equidae*. The concept *Attention* refers to the highly specific meaning of Attention in the field of psychology. The indexer falsely attributes any appearance of the word in its general meaning to this concept.

These examples illustrate some of the problems that can arise in the course of automatic indexing and it also shows that our method is an adequate means to identify them.

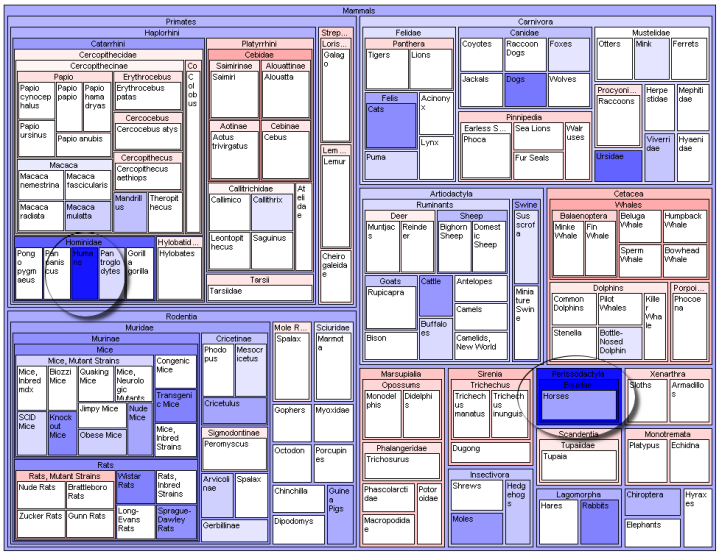


Figure 11 MeSH $IC_a - IIC$ Analysis of the node representing *Mammals*

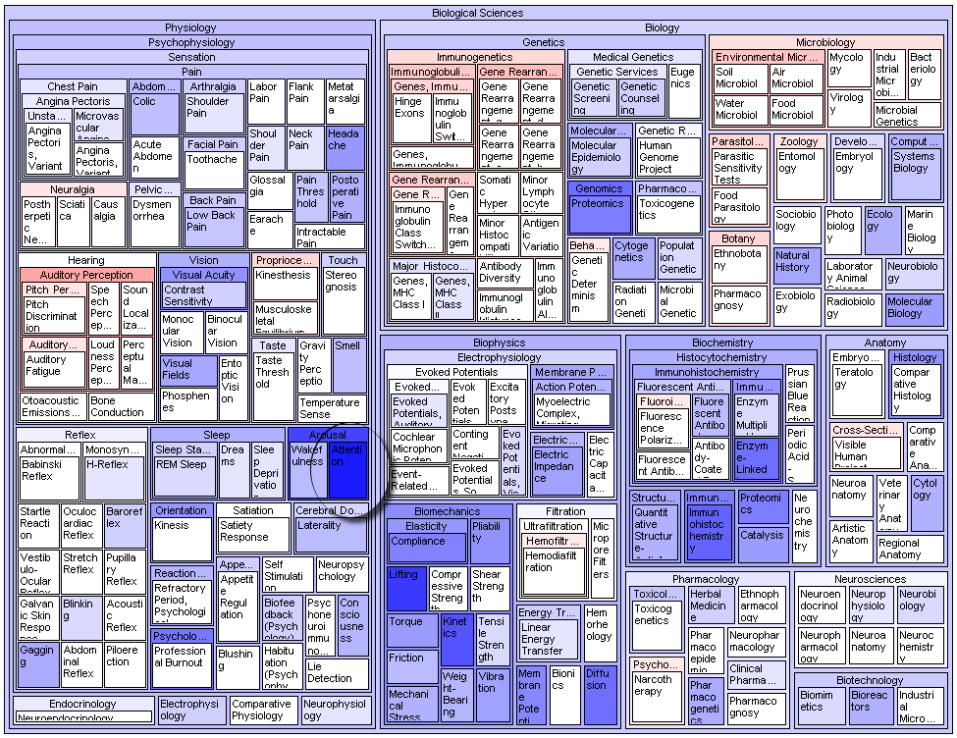


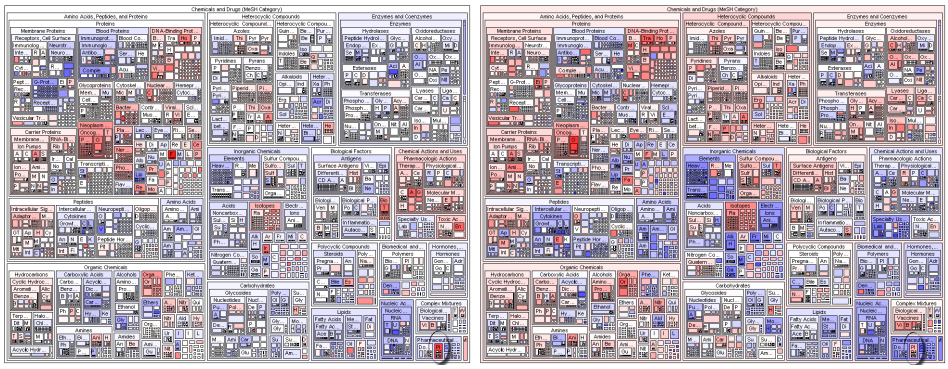
Figure 12 MeSH IC Difference Analysis ($IC_a - IIC$) of Biological Sciences

4.3 Comparing Automatic to Manual Annotations

When developing automatic indexing processes, usually a document set already annotated by a human expert is used, so that the two indexing results can be compared. While

one of the key points of our tool is that this is not strictly necessary, a direct comparison of selected keywords and found concepts is of course possible.

In this case, the analysis described in Equation 5 directly compares the two information contents of a thesaurus node and uses the same coloring conventions as the analyses before. Deeply tinted tiles still indicate problematic concepts, but in this analysis, the only reason can be errors in the automatic indexer.



(a) $IC_a - IC_m$ (b) $IC_a - IIC$

Figure 13 MeSH IC Difference Analysis of Chemicals

As less tiles are colored than in the analyses before, finding concepts that require a second look is much easier. Figure 13(a) illustrates this. The highlighted concept is *Plant Extracts*. It is used more often by manual annotators, probably because of lacking synonyms, as the preferred term ”Plant Extract” is rather abstract and not used very often in document abstracts. Compared to the analysis with IIC as reference (Figure 13(b)), the concept is more visible and sticks out of its vicinity.

5 Evaluation of the Automatic Indexing

As mentioned in the introduction, the goal of our method is to improve the quality of automatic indexing results by supporting human experts in the process of tuning the annotation system to the given task. In this section, we show that our method meets this goal. We use the annotation task described in section 2 as a basis and provide a quantitative evaluation of the annotations as well as the improvements achieved based on an assessment of the annotation result.

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. So we generally expect a higher recall than precision.

Beside the traditional Precision and Recall (with a binary decision, if a concept is regarded as a hit for a given keyword), we also perform a Generalized Precision and Recall evaluation, which is presented in the next section.

5.1 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 concepts that are not in the common term set Hahn and Schnattinger (1998); Keklinen and Jrvelin (2002); Maynard (2005); Maynard et al. (2006). The most recent proposal for generalizing precision and recall has been made by Euzenat Euzenat (2007), who defines generalized precision and recall in the following way:

$$Prec_\omega(A,R) = \frac{\omega(A,R)}{|A|} \; Rec_\omega(A,R) = \frac{\omega(A,R)}{|R|}. \tag{7}$$

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 approaches for semantic similarity measures including purely structural measures proposed by Leacock and Chodorow (1998) and Wu and Palmer (1994) as well as measures that are based on information theoretical concepts proposed by Resnik (1995), Lin (1998) and Jiang and Conrath (1997). A measure that tries to combine the two is proposed by Seco et al. (2004) 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 (1998) with Intrinsic Information Content as a basis for our experiments.

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

$$Similarity_L(c_1,c_2) = \frac{2 \cdot IIC(LCS)}{IIC(c_1) + IIC(c_2)} \tag{8}$$

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

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

$$Recall = \frac{\sum_{r \in R_d} \mathtt{max}_{a \in A_d} \; Similarity_L(r,a)}{|R_d|} \tag{9}$$

$$Precision = \frac{\sum_{a \in A_d} \mathtt{max}_{r \in R_d} \; Similarity_L(a,r)}{|A_d|}. \tag{10}$$

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.

5.2 Results

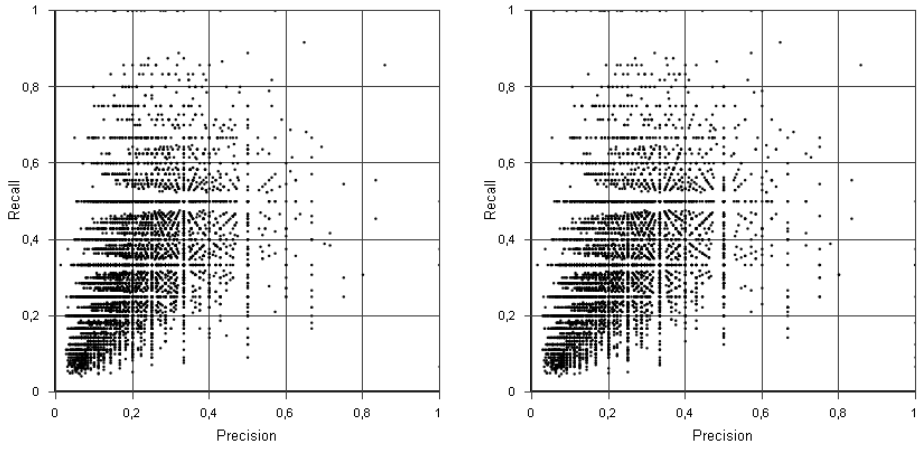
We improved the MeSH thesaurus for automatic indexing by identifying problematic concepts that are not suitable for our annotation system. As shown in the examples in Section 4, these are especially concepts with ambiguous meanings and concepts that are problematic for the involved normalizer. For the identification of these concepts, one person (familiar with our prototype, but no domain expert) needed about one hour. The result was a list of 18 concepts, containing for example the publication types (as they are not used for content description by the human annotators), the World Health Organization (WHO, falsely annotated for the occurrence of "who") and several psychological concepts like Attention, Inhibition, Identification, Set, Role and Aptitude (common terms as synonyms that are ambiguous with respect to the context).

Binary Precision and Recall. Table 3 presents the results for the traditional Precision and Recall measure. The numbers given correspond to the overall precision and recall over the complete document set. The absolute values are rather small, as only exact hits are mentioned. But we are especially interested in the question, if we can increase the results significantly.

The binary precision and recall analysis shows that the recall remains nearly unchanged by our improvements, as expected. On the other side, the precision is increased by about 2 percentage points, which is an improvement of 10.2%. Figure 14 provides a more detailed view of the results in terms of the precision and recall of individual documents in the set.

Set	# Keywords	∑Similarity	Recall	Improvement
Original	82319	26143	0.318	
Improved	82264	26105	0.317	-0.1%
Set	# Concepts	∑Similarity	Precision	Improvement
Original	133911	26143	0.195	
Improved	121396	26105	0.215	+10.2%

Table 3 Precision and Recall (Binary), 7376 Medline abstracts



(a) Original (b) Improved

Figure 14 Precision and Recall (Binary)

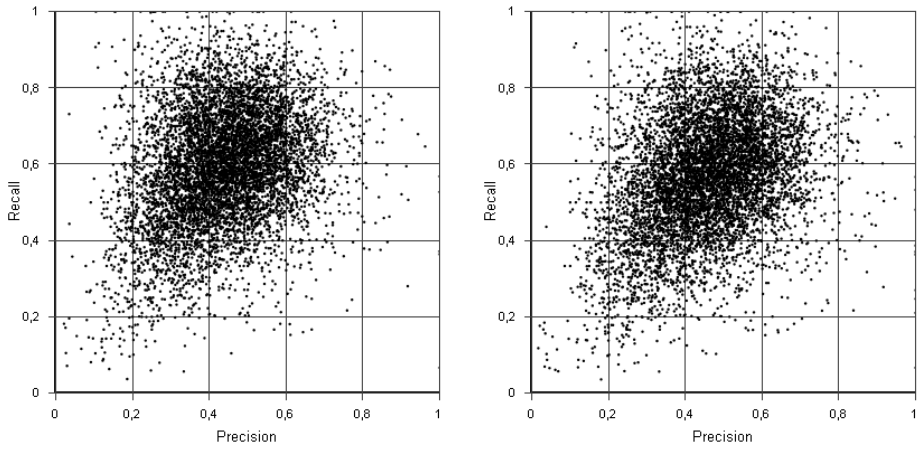
Generalized Precision and Recall. The results of the evaluation with Generalized Precision and Recall are presented by Table 4 and Figure 15. The precision is increased by 2.4 percentage points (or 5.7%). On the other hand, the recall is decreased by 1.6 percentage points (or 2.7%).

The Generalized Precision and Recall approach always produces higher values for Precision and Recall than the traditional one. This is one reason for the smaller relative improvement. But especially the significantly decreased recall indicates that the chosen similarity measure might be too tolerant. We only marked concepts that were only rarely used for manual annotations if at all. This means that the erroneous annotations in the original set contributed positively to the recall measure and thus the improvements led to a decrease.

This shows that the similarity measure has to be chosen very carefully. This remains a very complex task and should be an object of future research.

Set	# Keywords	∑Similarity	Recall	Improvement
Original	82319	48276	0.586	
Improved	82264	46938	0.571	-2.7%
Set	# Concepts	∑Similarity	Precision	Improvement
Original	133911	56660	0.423	
Improved	121396	54289	0.447	+5.7%

Table 4 Generalized Precision and Recall (Lin Intrinsic), 7376 Medline abstracts



(a) Original (b) Improved

Figure 15 Generalized Precision and Recall

6 Discussion and Conclusions

In this article 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 demonstrated some examples that show how a human expert can use our approach to supervise an automatic annotating system and thus keeps track of the overall quality of the information retrieval system. We also presented results of initial experiments in improving annotations based on the result of the analysis in which fixing some problematic concepts 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. In our experiments we revealed a substantial weakness of the generalized precision and recall approach. Nevertheless, we believe that this approach is necessary to adequately compare sets of annotations. Probably, a more restrictive similarity measure has to be chosen to avoid improvements of the results by accident. 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. In our improvements, we marked concepts that were not suitable for our annotation system. We propose this as a general way to improve existing thesauri for automatic annotation systems. For example, it could be noted that a given concept should only be used if the indexer can perform a correct noun phrase detection or context disambiguation and that some terms should only be considered, if a sophisticated disambiguation step is involved.

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 article, as well as an animated presentation of the interactive tool.

References

Bederson B.B., Shneiderman B. and Wattenberg M. (2002) 'Ordered and quantum treemaps: Making effective use of 2D space to display hierarchies', *ACM Trans. Graph.*, 21(4), pp. 833–854.

Burkart M. (2004) *Grundlagen der praktischen Dokumentation und Information, Band 1*, chap. Thesaurus, pp. 151–153, Saur.

Calmet J. and Daemi A. (2004a) 'Assessing Conflicts in Ontologies', Tech. rep., IAKS Calmet, University Karlsruhe (TH), Germany.

Calmet J. and Daemi A. (2004b) 'From entropy to ontology', Tech. rep., Institute for Algorithms and Cognitive Systems (IAKS), University of Karlsruhe (TH), Germany.

Clifford Gay M.K. and Aronson A. (2005) 'Semi-Automatic Indexing of Full Text Biomedical Articles', in 'Proceedings of the Annual Symposium of AI in medicien, AIMA 2005', .

Euzenat J. (2007) 'Semantic Precision and Recall for Ontology Alignment Evaluation', in 'IJCAI', pp. 348–353.

Ferber R. (1997) 'Automated Indexing with Thesaurus Descriptors: A Co-occurence Based Approach to Multilingual Retrieval', in 'Proceedings of the First European Conference on Research and Advanced Technology for Digital Libraries', vol. 1324 of *Lecture Notes In Computer Science*, pp. 233 – 252.

Hahn U. and Schnattinger K. (1998) 'Towards Text Knowledge Engineering', in 'AAAI/IAAI', pp. 524–531.

Jiang J.J. and Conrath D.W. (1997) 'Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy', in 'Natural Language Engineering - Proceedings of International Conference Research on Computational Linguistics (ROCLING X)', vol. 11, pp. 81–111, Tapei, Taiwan.

Keklinen J. and Jrvelin K. (2002) 'Using graded relevance assessments in IR evaluation', *J. Am. Soc. Inf. Sci. Technol.*, 53(13), pp. 1120–1129.

Leacock C. and Chodorow M. (1998) 'Combining Local Context and WordNet Similarity for Word Sense Identification', in C. Fellbaum (ed.), 'WordNet: An Electronic Lexical Database', chap. 11, pp. 265–283, MIT Press.

Lin D. (1998) 'An information-theoretic definition of similarity', in 'Proc. 15th International Conf. on Machine Learning', pp. 296–304, Morgan Kaufmann, San Francisco, CA.

Maynard D. (2005) 'Benchmarking ontology-based annotation tools for the Semantic Web', in 'UK e-Science Programme All Hands Meeting (AHM2005) Workshop "Text Mining, e-Research and Grid-enabled Language Technology"', Nottingham, UK.

Maynard D., Peters W. and Li Y. (2006) 'Metrics for Evaluation of Ontology-based Information Extraction', in 'WWW 2006 Workshop on "Evaluation of Ontologies for the Web" (EON)', Edinburgh, Scotland.

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

Seco N., Veale T. and Hayes J. (2004) 'An Intrinsic Information Content Metric for Semantic Similarity in WordNet', in 'Proceedings of the 16th European Conference on Artificial Intelligence', pp. 1089–1090, Valencia, Spain.

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

Shneiderman B. and Wattenberg M. (2001) 'Ordered Treemap Layouts', Online, ftp://ftp.cs.umd.edu/pub/hcil/Reports-Abstracts-Bibliography/2001-06html/2001-06.htm, visited 2007-02-19.

Smith M. and Fiore A. (2001) 'Visualization components for persistent conversations', in 'Proceedings of the SIG-CHI on Human factors in computing systems', pp. 136–143.

Stuckenschmidt H., de Waard A., Bhogal R., Fluit C., Kampman A., van Buel J., van Mulligen E., Broekstra J., Crowlesmith I., van Harmelen F. and Scerri T. (2004) 'Exploring Large Document Repositories with RDF Technology - The DOPE Project', *IEEE Intelligent Systems, special Issue on the Semantic Web Challenge*.

Wu Z. and Palmer M. (1994) 'Verb semantics and lexical selection', in '32nd. Annual Meeting of the Association for Computational Linguistics', pp. 133–138, New Mexico State University, Las Cruces, New Mexico.