Analysis of Structural Relationships for Hierarchical Cluster Labeling

Markus Muhr
Know-Center Graz
Inffeldgasse 21a
8010 Graz, Austria
mmuhr@know-center.at

Roman Kern
Know-Center Graz
Inffeldgasse 21a
8010 Graz, Austria
rkern@know-center.at

Michael Granitzer
Know-Center Graz
Graz University of Technology
Inffeldgasse 21a
8010 Graz, Austria
mgrani@know-center.at

ABSTRACT

Cluster label quality is crucial for browsing topic hierarchies obtained via document clustering. Intuitively, the hierarchical structure should influence the labeling accuracy. However, most labeling algorithms ignore such structural properties and therefore, the impact of hierarchical structures on the labeling accuracy is yet unclear. In our work we integrate hierarchical information, i.e. sibling and parent-child relations, in the cluster labeling process. We adapt standard labeling approaches, namely Maximum Term Frequency, Jensen-Shannon Divergence, χ^2 Test, and Information Gain, to take use of those relationships and evaluate their impact on 4 different datasets, namely the Open Directory Project, Wikipedia, TREC Ohsumed and the CLEF IP European Patent dataset. We show, that hierarchical relationships can be exploited to increase labeling accuracy especially on high-level nodes.

Categories and Subject Descriptors

H.3.1 [**Content Analysis and Indexing**]: Linguistic processing; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Algorithms

Keywords

Cluster Labeling, Statistical Methods, Topic Hierarchies, Structural Information

1. INTRODUCTION

Browsing large-scale document collections usually requires a structural organization form like topic hierarchies. Unsupervised machine learning techniques, foremost document clustering, overcome the labor intensive, manual creation of such topic hierarchies by automatic partitioning of unstructured document collections into browse-able cluster hierarchies. This cluster based browsing approach has been shown to successfully improve access to unstructured document collections [5, 16].

However, even if an algorithm achieves a perfect hierarchical partitioning, users must guess the content of each cluster somehow. Automatically generated labels provide such a descriptive cluster summary. Obviously, the label quality strongly influences the navigation effectiveness as shown by human created topic hierarchies like the Open Directory Project: while users may disagree on the exact label of a topic, every user can exploit labeling information for navigation purposes as long as labeling is of high quality - this is especially true for high level nodes like sports, computers etc.

Most existing labeling approaches extract labels by comparing term distributions of a cluster to a reference collection and taking the statistically most discriminative terms. Intuitively, for a flat partitioning this seems to be sufficient, but insufficient for creating topic hierarchies similar to the Open Directory Project (ODP) [1]; Child clusters have to be described in the context of their parent cluster and must not contain the same labels. Such a constraint cannot be ensured without taking structural relationships between clusters into account. Moreover, label quality tends to decrease on higher levels due to higher degree of abstraction. Most state of the art labeling approaches (e.g. [11, 1]) do not use structural relationships. Although there are approaches considering hierarchical relationships - either through supervised learning [19] or through hierarchical post-processing of flat cluster labels [14] - to our best knowledge there is no systematic investigation whether the intuitive claim above holds or not. Intuitively it also seems to be natural that labeling performance as well as the influence of structural relationships depends on a topic hierarchies structural properties - a claim hard to investigate due to missing test corpora.

In this paper we investigate the influence of hierarchical relationships on the cluster labeling process. We extend standard labeling approaches, namely Maximum Term Frequency, Jensen-Shannon Divergence, χ^2, and Information Gain, to include structural information. First, Maximum Term Frequency labeling is extended by a sibling based weighting scheme, yielding to a new labeling algorithm called ICWL (Inverse Cluster frequency Weighted) labeling. Second, we extend all labeling approaches with parent-

[1] `http://dmoz.org`

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
SIGIR'10, July 19–23, 2010, Geneva, Switzerland.
Copyright 2010 ACM 978-1-60558-896-4/10/07 ...$10.00.

child relationships. Comparing all labeling approaches on four datasets, namely the Open Directory Project (ODP), Wikipedia, TREC-Ohsumed and European Patents (EP), shows that hierarchical information influences the labeling process. Especially Wikipedia shows the biggest dependency on hierarchical information, followed by the ODP dataset. This finding yields to an important point for future work on this topic: top level nodes, which are most important in the users browsing process, are labeled badly. *Hence, the browsing process of automatically created cluster hierarchies could be improved by using hierarchical cluster label algorithms.*

With our work we *contribute* to the field of cluster labeling by

- extending standard labeling approaches to take use of hierarchical information
- showing, that sibling relationships can be exploited to improve statistical labeling methods
- showing, that the structure of the hierarchy and the domain of the test dataset have a strong influence on the labeling accuracy, especially for top level nodes, which are crucial for user navigation
- using traditional, but also new test corpora for evaluating cluster labeling algorithms

The paper is structured as follows: first a related work section provides an overview of the state of the art relevant for our work, second we introduce a formal definition of structure based labeling, followed by the third part, a description of the utilized corpora (ODP, Wikipedia, Ohsumed, European Patents) together with the implemented preprocessing steps. The paper ends by outlining results with an obligatory discussion pointing out our findings, and finally a summary of our work with a discussion of the implications for future work.

2. RELATED WORK

Standard labeling approaches extract most prominent terms in a specific cluster by statistical feature selection [11]. A straightforward feature selection method takes the maximum sum of the individual term frequencies of documents assigned to a cluster [5]. In [14], a weighting schema has been introduced to improve maximum sum of term frequencies by neglecting stop-words or general words.

However, maximum-sum approaches prefer terms which are over-represented in the whole document collection. This increases the probability that all cluster are getting similar labels. More sophisticated approaches consider a terms discriminative power compared to a reference collection - usually the documents of all sibling clusters. Well-known methods include an adapted versions of Information Gain [7], χ^2 Test [14], and the Jensen-Shannon Divergence (JSD) [2]. Comparative studies of reference collection based labeling approaches done in [1] favor JSD for the non-hierarchical case. However, in the hierarchical case such an in-depth comparison between labeling approaches is missing. Further, the evaluation conducted in [1] favors leaf nodes of a hierarchy without addressing questions on the influence of hierarchical structures.

Besides statistical term selection methods, researchers focused on using different document parts like title [5], hyperlink anchors [9] etc. or different features like named entities [17], frequent phrases [13] or text summarization [15]. However, our work focuses not on finding the very best labeling approach including the best document representation, but to include structural properties into statistical labeling approaches. Treeratpituk et. al. [19] addressed this problem via supervised learning. Weights for different term importance measures depending on parent-child relationships are estimated using supervised learning. However, their method needs training data to determine the actual weights. Through supervised training and the use of synonyms in the labeling process the actual influence of hierarchical relationships remains unclear. Further work in hierarchical labeling has been done by Popescul and Ungar [14], where parent child relationships are exploited in a post processing step and not directly included into the feature selection process. Hence, errors from the labeling process are propagated to the post processing step.

In the past years researchers investigated the usage of external knowledge to enhance machine learning tasks through extending the given term set with synonyms, hypernyms, hyponyms etc. Most prominent resources for this task are WordNet [3] and Wikipedia [1]. Especially Carmel et. al. [1] showed that using Wikipedia as a thesaurus and applying this thesaurus as post-processing step to statistical labeling approaches improves cluster labeling dramatically. Although they used statistical labeling approaches in a first step, hierarchical information was not included. So by improving the statistical labeling their approach may be further enhanced.

Evaluating cluster labeling approaches face the dataset sparsity problem. Most methods mentioned above use the Open Directory Project (ODP) [12] as well as the flat 20 newsgroup dataset. This basically restricts the obtained results on the domain of web resources. Whether the results apply also to other domains like patents, medicine etc. remains an open question. We address this question by conducting experiments, in addition to the ODP, also on Wikipedia [6], the TREC-Ohsumed collection [18] and the European Patents provided by the 2009 CLEF IP Task [4].

3. STRUCTURE BASED LABELING

In order to analyze the impact of hierarchical relationships we gradually incorporate structure information into well-known labeling techniques. We use a maximum term weight labeling approach as well as a reference collection based labeling approach to estimate the labels of a cluster. Through weighting methods based on the position of a document relative to the label candidate cluster, we introduce structure information in terms of (i) sibling relations and (ii) parent-child relations. All approaches consider bag-of-word document representations and hierarchical relationships among clusters. Before describing the labeling approaches in detail, we introduce a formalism for denoting cluster hierarchies, document sets as well as a notation for selecting particular sub-trees and and sub document sets.

Formally, we define $\mathbf{D}$ as the set of all documents and a cluster c_l as a sub set of documents $c_l \subseteq \mathbf{D}$; $\mathbf{C} = \{c_1 \dots c_k\}$ denotes as set of non-overlapping clusters, i.e. $\forall_{i \neq j} c_i \cap c_j = \emptyset$. The hierarchical structure between clusters is formalized as is-a relationship $c_j \rightarrow c_i$ ($c_j \Longrightarrow c_i$), defining that c_j is the parent (direct parent) of c_i. For specifying the set of clusters with parent c_j, we write $\mathbf{C}_{c_j \rightarrow *}$ and denote all documents contained in this sub-hierarchy as $\mathbf{D}_{c_j \rightarrow *}$. Se-

mantically the is-a relationship should resemble a classical topic hierarchy, which assumes that if a document is assigned to a topic, it is also assigned to its parent topic. For the rest of this paper we consider the is-a relationship as implicitly given when referring to the set of clusters $\mathbf{C}$ or any subset of them. Further, documents are represented as a term frequency vector $d \in \Re^d$. Using a vector representation allows us to consider different weighting methods to assess the a-priori importance of terms in a corpus resp. a cluster. Labels L_j of a cluster c_j are represented as set of terms $L_j = \{t_1 \dots t_l\}$.

3.1 Maximum Term Weight Labeling

Labeling can now be seen as function $L_j \leftarrow label(\mathbf{C}, c_j)$ selecting the most suitable terms for describing cluster c_j as label L_j. The simplest labeling functions are basic feature selection techniques [11]: Documents assigned to a cluster are aggregated and the k largest features are taken as labels. We refer to this as Maximum Term Weight Labeling (MTWL): given cluster c_j as labeling candidate the MTWL can be written as

$$L_j \leftarrow best_k\Big(\sum_{d_i \in \mathbf{D}_{c_j \rightarrow *}} d_i\Big) \tag{1}$$

where $best_k\Big(v\Big)$ is a function returning the terms associated with the k largest dimensions of vector v. In our experiments we refer to this approach as $MTWL_{raw}$.

Maximum Term Weight Labeling strongly depends on the document representation. Given only term frequency document vectors, labeling will likely extract terms occurring in a large number of documents with high frequency. Such labels do not necessarily discriminate between clusters. Global weighting schemes like TFIDF or Okapi BM25 [11] allow to increase the discrimination capability of labels based on the underlying document distribution. Since documents in the same cluster share similar terms, the inverse document frequency reduces the influence of terms occurring in a high number of documents and therewith in a high number of cluster. To introduce such global weighting, MTWL can be simply extended as follows:

$$L_j \leftarrow best_k\Big(\sum_{d_i \in \mathbf{D}_{c_j \rightarrow *}} idf_{global} \cdot tfWeight(d_i)\Big) \tag{2}$$

where $idf_{global} \in \Re^d$ is a vector containing the corpus dependent inverse document frequencies, $tfWeight()$ is a function applying the document specific part of the weighting scheme and $\cdot$ is the Hadamard point-wise product. The inverse document frequency for a term k is calculated as

$$idf_{global,k} = \log\Big(\frac{|\mathbf{D}|}{\#(t_k, \mathbf{D})} + 1\Big) \tag{3}$$

with $\#(t_k, \mathbf{D})$ returning the number of documents in the collection containing term k. For the term frequencies we used the document specific part of the standard Okapi BM25 as well as plain term frequencies[2]. Since idf_{global} is defined over the collection $\mathbf{D}$ of all documents we refer to this approach as global weighting approach.

[2]We split the weighting scheme in a collection and a document specific part in order to have a homogeneous notation over our different labeling approaches.

Global weighting penalizes terms, which are over-represented in the whole collection. However, terms over-represented in a particular cluster sub-tree only, will be likely selected for all siblings in the cluster hierarchy. For example, given that term t_k is over-represented in cluster c_j and equally distributed among the direct children c_i with $c_j \Longrightarrow c_i$, then it is very likely that term t_k will become a label of the direct child c_i. Hence, term distributions among siblings have to be taken into account to avoid siblings getting similar labels.

For considering sub-tree dependent term distribution we again add a local, sub-tree based inverse document frequency term to eq. 2. Formally, the labeling function is defined as

$$L_j \leftarrow best_k\Big(\sum_{d_i \in \mathbf{D}_{c_j \rightarrow *}} idf_{global} \cdot idf_{local,j} \cdot tfWeight(d_i)\Big) \tag{4}$$

where $idf_{local,j}$ is the inverse document frequency vector over the document collection $\mathbf{D}_{c_p \rightarrow *} with\ c_p \Longrightarrow c_j$ where c_p is defined as the parent cluster of c_j. Simply speaking this document collection consists of all documents in the subtree spanned by the parent cluster c_p. In particular the idf entry for term k in cluster c_j is calculated as

$$idf_{local,j} = \log\Big(\frac{|\mathbf{D}_{c_p \rightarrow *}|}{\#(t_k, \mathbf{D}_{c_p \rightarrow *})} + 1\Big) \tag{5}$$

with $\#(t_k, \mathbf{D}_{c_p \rightarrow *})$ returning the number of documents in the reference collection which contain term k. In our experiments we refer to this approach as $MTWL_{idf}$.

3.2 Reference Collection based Labeling

Besides taking the largest dimension of a centroid vector, comparative statistics like the χ^2-Test or the Jensen-Shannon Divergence (JSD) can estimate whether occurrences of a term differ between a cluster and a reference collection with statistical significance. Such terms yield good labels for a cluster. In our Reference Collection based Labeling approach (RCL) we use well known comparative statistics, namely the Jensen-Shannon Divergence, Information Gain and χ^2, in order to compare terms contained in clusters to terms contained in a reference collection of documents, denoted as $\mathbf{D}_{ref}$. The k terms with the best test values are taken as labels. Hierarchical information is incorporated through the selection of the reference collection. Formally, we denote

$$L_j \leftarrow best_k\Big(JSD(\mathbf{D}_{ref}, \mathbf{D}_{c_j \rightarrow *})\Big) \tag{6}$$

as labeling function where $JSD(\mathbf{D}_{ref}, \mathbf{D}_{c_j \rightarrow *}) \rightarrow \Re^d$ returns the Jensen-Shannon Divergence (see [2]) for each dimension in form of a d-dimensional vector. Again, $best_k\Big(v\Big)$ is used to select the k terms with the best statistical test value. Similarly we abbreviate the Information Gain as $IG(\mathbf{D}_{ref}, \mathbf{D}_{c_j \rightarrow *})$ [7] and χ^2 as $\chi^2(\mathbf{D}_{ref}, \mathbf{D}_{c_j \rightarrow *})$ [14]. The probability for a term is estimated in a standard manner as the number of occurrences of a term divided by the total number of occurrences of terms.

For labeling cluster c_j we define the reference collection as all documents belonging to the cluster sub-tree of its direct parent excluding all documents contained in c_j. Formally, the reference collection is given as $\mathbf{D}_{ref} = \mathbf{D}_{c_p \rightarrow *} \setminus \mathbf{D}_{c_j \rightarrow *}\ with\ c_p \Longrightarrow c_j$, where c_p is the parent of c_j. In the non-hierarchical case this corresponds to the best known standard labeling approaches [1].

3.3 Inverse Cluster Weight Labeling

MTWL incorporates hierarchical information through a specialized weighting function $idf_{local,j}$. The weighting function depends on the term distribution over documents in the sub-tree, but does not take the term distribution over sibling cluster into account.

To integrate sibling information we add another weighting factor. The weighting factor is inspired by the recently introduced Class-Feature-Centroid (CFC) classifier using class discriminative terms for achieving classification accuracies similar to Support Vector Machines [10]. Similarly we want to take use of clusters discriminative terms. Basically, if one term occurs often in one sibling cluster only, this term should be preferred over terms occurring in all sibling cluster: a term k in cluster c_j is weighted by its inverse cluster frequency calculated as

$$icf_{j,k} = \exp\Big(\frac{\#(t_k, \mathbf{D}_{c_j \to *})}{|\mathbf{D}_{c_j \to *}|}\Big) \log\Big(\frac{\#(c_p)}{\#(t_k, c_p)} + 1\Big) \qquad (7)$$

with c_p being the direct parent of c_j, $\#(t_k, c_p)$ being the number of direct subcluster of c_p containing term k and $\#(c_p)$ being the number of direct sub-clusters. The exponential component, similar to the CFC classifier, promotes terms occurring in a larger fraction of documents. In our experiments we refer to this approach as $ICWL_{raw}$ if $MTWL_{raw}$ is extended by the inverse cluster frequency weight and to $ICWL_{idf}$ in the case of $MTWL_{idf}$.

3.4 Hierarchical Labeling

Especially in the hierarchical case, relying only on sibling information may be problematic; a term occurring often in the parent cluster (and all its documents) may also occur often in one or several child clusters. While incorporating sibling information potentially removes parent labels equally distributed over potentially all siblings, it cannot overcome parent labels occurring often in a few child clusters. Hence, parent child relationships have to be taken into account.

Hierarchical labeling extends all labeling approaches introduced before, by weighting the influence of a term inverse proportional to the path length of the child cluster to the label candidate cluster c_j. By denoting the path length between two clusters as $l(j,i)$, the labeling function of cluster c_j can be formally written as

$$L_j \leftarrow best_k\Big(\sum_{c_i \in \mathbf{C}_{c_j \to *}} \frac{1}{l(j,i)} * cf_{l(j,i)} \cdot v_{j,i}\Big) \qquad (8)$$

where $cf_{l(j,i)}$ is the sibling based cluster frequency term vector and $v_{j,i}$ is the result of the comparison statistics formally written as

$$v_{j,i} = idf_{global} \cdot idf_{local,j} \cdot \sum_{d_o \in c_i} tfWeight(d_o) \qquad (9)$$

in case of MTWL labeling strategy. Simply speaking, $v_{j,i}$ is the centroid vector of cluster c_i weighted in the local context of the label candidate cluster c_j. This principle is similarly applied to the ICWL and the RCL approach.

Contrary to the inverse cluster frequency weighting above, the cluster frequency term vector penalizes terms occurring only in a single cluster on a particular hierarchy level (i.e. $l(j,i)$). The cluster frequency weight for term t_k is simply the number of clusters a term occurs in divided by the total number of clusters on this particular hierarchy level. The idea behind the cluster frequency weight is to promote terms occurring in a higher number of child clusters since those terms are most likely representative labels for their parent. In our experiments, all results involving hierarchical weighting are prefixed with "Hier".

4. DATASETS

For our experiments we used 4 different datasets: two general domain corpora, namely Wikipedia and Open Directory Project, and two domain specific corpora, namely European Patents (EP) and Ohsumed. All datasets have been preprocessed in the same way: document tokenization has been done using OpenNLP[3]; Tokens haven been stemmed afterwards using the Snowball[4] stemmer. Finally, stop-words have been removed by using the list supplied by the Snowball stemmer.

Open Directory Project (ODP): We imported a large part of the hierarchy including the top categories arts, business, games, health, home, news, society, and sports with their complete subtree. We took only hard links into account ignoring symbolic links and related topic links. Letter categorizations are ignored as well. This yielded about 150,000 categories and about 800,000 documents. In order to compare the effect of the manually created descriptions and titles for each ODP entry, we created two sub-datasets. In the first dataset, named ODP Title & Description, each document consists of the description and title as provided in the ODP hierarchy. For the second dataset, named ODP HTML, we crawled the HTML page a ODP entry pointed to and all HTML pages links in the crawled page pointed to, i.e. we performed a crawl of depth 2. This crawling strategy should provide sufficient content rich pages producing a rather different dataset compared to the classical ODP Title & Description dataset.

Wikipedia: To extract the structural information out of the Wikipedia we started with the XML dump of the English version. From each entry within the dump we extracted the title and all links that indicate an assignment to a Wikipedia category. This was done for all articles and category pages such that we were able to reconstruct the classification relationships. We filtered out categories that do not carry any semantic information and assembled a category blacklist (e.g. "Wikipedia maintenance') and filtered out categories containing the word "by", "of" and "in" to eliminate categories like "Authors by Year". In order to create a tree structure of of the acyclic category graph we started by each main topic - namely arts, computing, health, and sports - we traversed the graph in a breath-first manner with a maximum depth of 10. Since a breath-first search of depth 10 would return a too large portion of the Wikipedia graph, we randomly chose 10 outgoing links and 80 documents for each topic. Roughly we had about 50,000 categories with about 400,000 documents as test dataset.

TREC Ohsumed: We used the Ohsumed collection from the 2001 TREC evaluation. The hierarchical structure has been obtained downloading the Mesh Tree hierarchy[5] of 2004 with 7724 different categories and 348,564 documents.

European Patents (EP): European patents are taken from

[3] `http://opennlp.sourceforge.net/`
[4] `http://snowball.tartarus.org/`
[5] `http://www.nlm.nih.gov/mesh/`

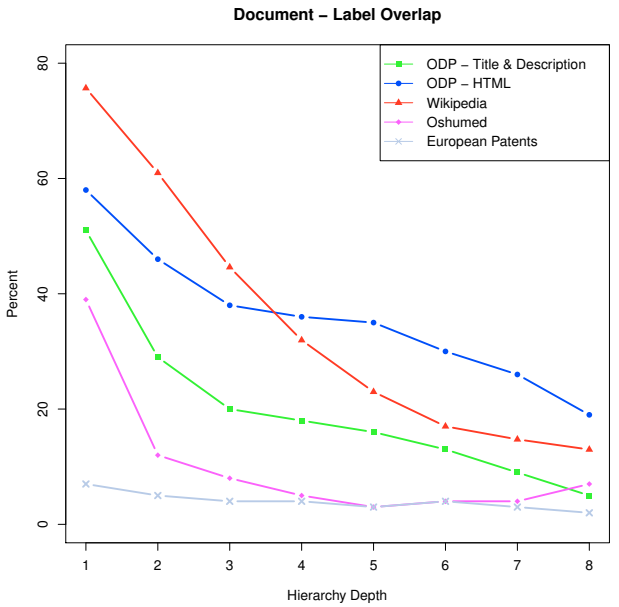

Figure 1: Document-Label-Overlap: Fraction of documents containing all label terms depending on the distance to the category.

the dataset which has been created for the Intellectual Property track of CLEF 2009 (CLEF-IP) [8]. From this dataset we selected only patents that were granted and limited the timespan from 1991 to 2000. We ended up with 265,409 patents, each of them having at least one assignment to the IPC classification scheme[6]. This IPC classification hierarchy consists of over 60,000 classes arranged in a tree-like manner with 8 root categories. The claims section has been used as document content.

5. RESULTS

In order to measure labeling accuracy we use the mean average precision (MAP) averaged over all categories. To calculate the MAP, category labels - similar to documents - are tokenized, stemmed and stopword filtered resulting in a set of terms. This set of terms is compared to the ranked list of candidate terms returned by the labeling algorithm, which gives the MAP value for one category.

We did not use any synonyms or external linguistic resources. To ensure that terms in the document set contain the terms extracted from the category labels, we estimated the Document-Label-Overlap, as outlined in the next section.

5.1 Document Label Overlap

The Document-Label-Overlap estimates whether a certain label of a cluster is contained in its connected documents at all. The Overlap is calculated as the fraction of documents containing all label terms to the number of total documents in the sub-tree with depth d. Thus, the overlap determines the baseline on getting a correct label for a topic. Further, by considering the overlap of documents with a particular path length d to the label candidate cluster we get an evidence on the influence of documents on particular hierarchy depths.

Figure 1 depicts the document label overlap for all datasets. Clearly, results show a significant decrease in the fraction of documents containing the actual label with the hierarchy level; a correct label is more likely found in documents close to the cluster. Furthermore, documents with high path lengths are more specialized and thus tend to use a more specialized vocabulary. For example, an article on Support Vector Machines might not mention the words machine learning explicitly, since it is a specialized topic in the field of machine learning. Hence, this analysis supports the evidence that structural properties play a role in cluster labeling.

A comparison between datasets point out interesting differences: for Ohsumed, ODP with Title & Description and the Wikipedia dataset the overlap drops significantly with increasing depth while it decreases rather slowly for the ODP HTML dataset. Clearly, ODP HTML contains more terms per document therewith increasing the likelihood of finding the correct label. The European Patents dataset shows its special nature: the overlap is constantly low over all hierarchy depths.

5.2 Labeling Accuracy

To evaluate the influence of the hierarchy depth on the labeling process, we plot the MAP on each dataset and labeling approach for sub-hierarchies of depth 1-8. We limit our analysis to sub-hierarchies with a maximum depth of 8 since there are too few sub-hierarchies with a larger depth making a statistical evaluation infeasible. Note also that the distribution of hierarchies is skewed: there are far more hierarchies with depth 1 than with depth > 1. Therefore, the overall labeling accuracy is approximately the labeling accuracy of depth 1 sub-hierarchies. This is contrary to the browsing behavior of a user who needs high labeling accuracy on the top nodes, i.e. on sub-trees with depth >1.

Figure 2 shows the labeling accuracy for the ODP dataset, split into description based documents and crawled HTML based documents. Figure 3 reports results on the Ohsumed and Wikipedia dataset. Note that for the clarity of presentation we only show the JSD labeling approach for RCL based labeling techniques. Compared to IG and χ^2 (as well as their hierarchical counterparts), JSD always achieved the best performance. This supplements the findings in literature and extends them also to the hierarchical case, see [1].

Comparison of labeling techniques: Comparing the different labeling techniques it can be seen that the sibling based labeling approach with local and global weighting $ICWL_{idf}$ is performing about as good as the maximum term weighting approach, with exception of the the Wikipedia dataset where the integration of the sibling information does improve the accuracy. The labeling method that uses a reference collection JSD provides good results for every dataset. This is especially pronounced for the ODP Title & Description dataset when incorporating the hierarchical structure into the creation of the reference collection.

Flat vs. Hierarchical labeling: Table 1 depicts the absolute MAP differences between hierarchical and flat approaches. For each of the tested combinations of datasets and labeling algorithms, the integration of hierarchical information always improves the accuracy. One exception is the Ohsumed dataset, where hierarchical and flat methods perform ap-

[6] http://www.wipo.int/classifications/ipc/en/

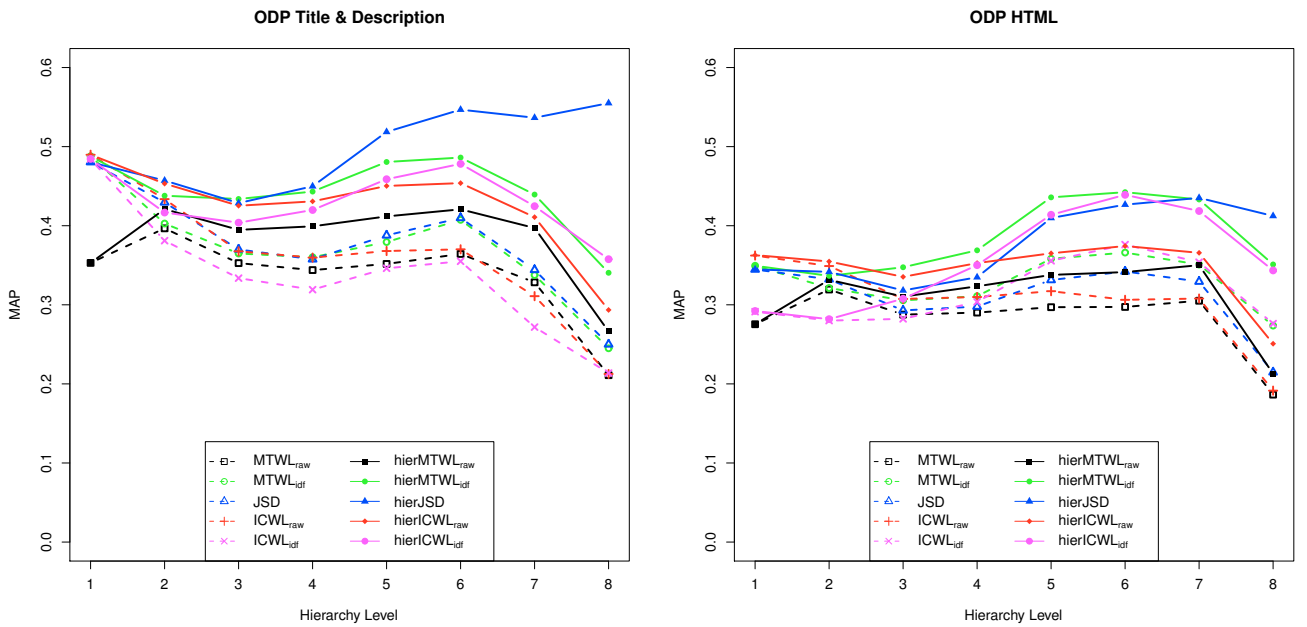

Figure 2: Performance of the different labeling algorithms for the Open Directory Project dataset. Algorithms that exploit the hierarchical structure generally produce better results.

proximately equal. Analyzing the sub-trees of top nodes we could observe a high variance of the influence of hierarchical labeling approaches which deserves further analysis. We assume that the increase depends on structural properties. A assumption to be validated in future work.

Ohsumed comes from a rather narrow domain compared to ODP and Wikipedia and is structurally different from all other datasets: the MESH-categorization includes documents only at leaf categories. Further, descriptions are rather small even compared to the ones from ODP. One can expect that these descriptions will not contain information about categories despite their direct parent - a fact also supported by the Document-Label-Overlap. For this reason, it is not surprising that hierarchical information increases labeling accuracy only slightly, since an already sparse information on term distributions is further reduced. This is also reflected by the fact that no labeling algorithm could outperform the other. The relative average difference for the MAP measure of the hierarchical labeling approaches in comparison with their flat counterparts is 0.004 for this dataset.

The *Wikipedia* dataset and the *ODP datasets* show a completely different picture. Flat labeling accuracy decreases significantly with the depth of the hierarchy; hierarchical labeling gives a slight average accuracy increase of 0.059 in case of the ODP HTML dataset, an accuracy increase of around 0.099 MAP on average in case of the ODP Description & Title dataset and a larger accuracy increase of around 0.132 MAP on average in case of Wikipedia (see also table 1). In case of the ODP dataset, the increase through hierarchical relationships on the Title & Description dataset is slightly better than on the HTML dataset. As depicted by the Document-Label-Overlap analysis, the HTML crawled documents most likely contain a broader range of terms compared to the title and description of the original ODP dataset. A broader range of terms increases the likelihood that a document contains a cluster label, even if it is farther away. This indicates that non-specialized documents on deeper nodes in a tree do not decrease labeling accuracy - a rather seldom case in topic hierarchies. In such a case, hierarchical weighting actually removes information on the topic instead of reducing the impact of certain more specific documents.

Regarding *Wikipedia*, the graph like category structure has to be considered, which has been adopted to a tree-structure in our case. For this reason, we took four main categories (Arts, Computing, Health, Sports) and created a topic hierarchy using the mentioned sampling strategy. While the sampling strategy seems to be fair, it does not reproduce the correct Wikipedia Category graph. Instead, we get a rather balanced tree like structure. Nevertheless, given such a balanced structure it is quite obvious that local idf weighted as well as hierarchical approaches benefit to a large degree. Especially the labeling of high level nodes could be dramatically increased through incorporating hierarchical information and significantly outperforms the non-hierarchical approaches. Moreover, this holds for all different labeling approaches.

The results for the *European Patents* are not depicted due to rather low accuracies. Nevertheless we mention them to support recent findings in the CLEF-IP challenge where it was shown that patents are a rather specialized domain [8]. Well-known information retrieval approaches failed to achieve good results in the challenge and it seems to be the same with cluster labeling methods. All approaches fail completely by only achieving values in the field of 0.03, although it seems to be the case the ICWL again improved the results. Also, the Document-Label Overlap shows that

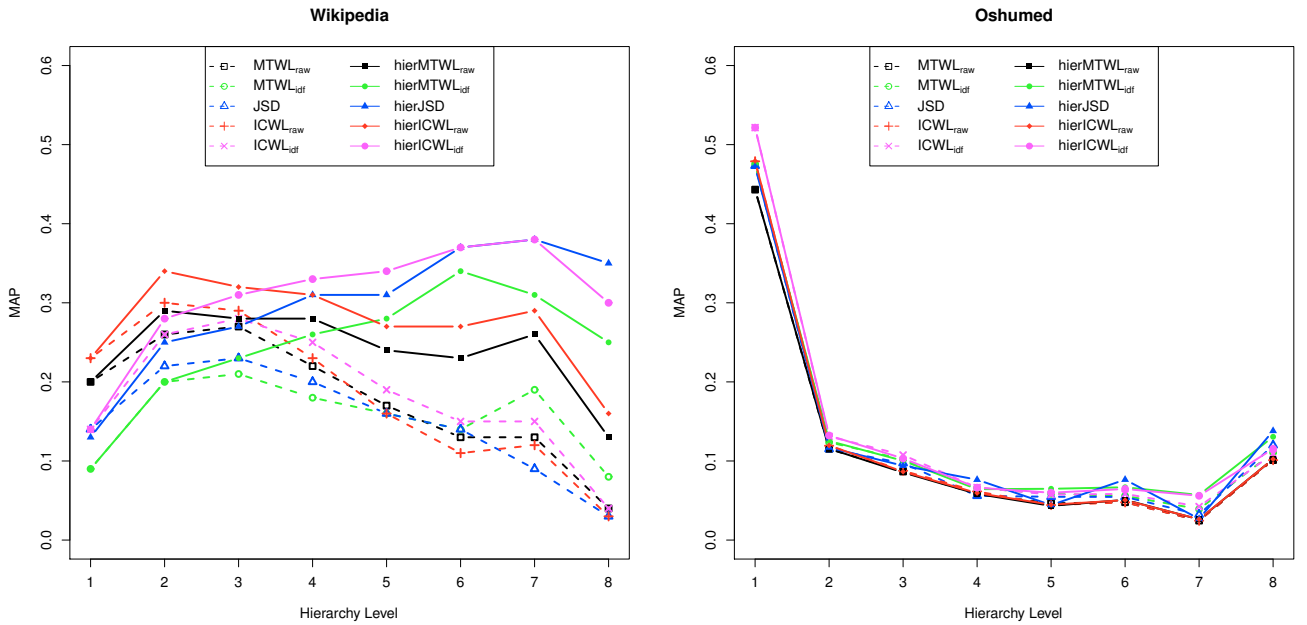

Figure 3: Performance of the different labeling algorithms for the Wikipedia and the Oshumed dataset. Both datasets demonstrate different characteristics, for the Wikipedia dataset the structural information increases the labeling performance, whereas for the Oshumed dataset neither the sibling nor the child-parent relationships help to find matching labels.

	$MTWL_{raw}$	$MTWL_{idf}$	JSD	$ICWL_{raw}$	$ICWL_{idf}$	Average
ODP - Title & Description	0.06	0.09	0.15	0.08	0.12	0.099
ODP - HTML	0.04	0.07	0.09	0.05	0.05	0.059
Wikipedia	0.08	0.12	0.19	0.12	0.16	0.132
Oshumed	0.00	0.01	0.01	0.00	0.00	0.004

Table 1: Average relative difference of the MAP for all hierarchy levels greater than 2 for all datasets between the different methods either with and without exploitation of hierarchical information. Exploiting the hierarchical structure always improves the accuracy, although for the Oshumed dataset the difference is not pronounced.

only below 10 % of the documents connected to a category contain a label term.

Overall our results imply that incorporating hierarchical information improves labeling accuracy on average.

Moreover, our results have impact on the evaluation of cluster labeling approaches for browsing topic hierarchies. Viewed from a users point of view, flat labeling approaches support the browsing of leaf nodes rather than the browsing of high level nodes - a result quite contradictory to the users need. Especially in the case of the de-facto standard benchmark dataset, the ODP Description & Title dataset, this has to be taken into account for future evaluations. First, sampling data for an evaluation should consider the a-priori distribution of sub-hierarchies with different depths. By using depth independent random samples for cluster labeling evaluation it is very likely to draw hierarchies of depth 1 and to achieve good labeling performance using flat labeling approaches. Second, hierarchies of different depth should be evaluated separately in order to deduce the impact of the labeling strategy on the users navigational support.

6. CONCLUSION AND OUTLOOK

Our results show that structural relationships influence the labeling accuracy. Using sibling information increases labeling accuracy in some datasets; integrating hierarchical information produces better labeling results for all datasets. This insight has several consequences.

Firstly, evaluation of cluster labeling approaches have to take hierarchical properties into account, especially if the goal is to support user navigation.

Secondly, correlations between the properties of a hierarchy, like for example maximum depth, branching factor, documents per leaf node etc., the richness of the assigned documents and the achievable labeling accuracy should be further analyzed. While we followed the evaluation approach conducted by other researchers in the field, there should be a closer evaluation whether cluster hierarchies and manually created hierarchies resemble the same statistical properties w.r.t the document collection.

Thirdly, more sophisticated approaches like for example the extension of JSD with hierarchical information may further increase the accuracy in the hierarchical case. Although we integrated parent-child relationships in an ad-hoc manner, we observed an effect on the labeling accuracy. Clearly we would expect more sophisticated approaches to increase accuracy further.

Fourthly, labeling accuracy is strongly domain dependent. The generalization of labeling approaches to different domains remains an open issue.

Finally, external knowledge in form of thesauri, ontologies etc. has to be considered also in the hierarchical case. We restricted our work to term frequency vectors only and focused poorly on statistical approaches that do not incorporate any external knowledge in form of thesauri, ontologies etc. However, our labels depend solely on the document representation and hence the term frequency vectors may be replaced by more sophisticated preprocessing utilizing external knowledge. Furthermore, labeling approaches using external knowledge most often depend on good statistical label selection and thus our approach contributes to their improvement.

Acknowledgments

The Know-Center GmbH Graz is funded within the Austrian COMET Program - Competence Centers for Excellent Technologies - under the auspices of the Austrian Federal Ministry of Transport, Innovation and Technology, the Austrian Federal Ministry of Economy, Family and Youth and by the State of Styria. COMET is managed by the Austrian Research Promotion Agency FFG.

7. REFERENCES

[1] D. Carmel, H. Roitman, and N. Zwerdling. Enhancing cluster labeling using wikipedia. In *SIGIR '09: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 139–146. ACM Press, 2009.

[2] D. Carmel, E. Yom-Tov, A. Darlow, and D. Pelleg. What makes a query difficult? In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 390–397. ACM Press, 2006.

[3] O. S. Chin, N. Kulathuramaiyer, and A. W. Yeo. Automatic discovery of concepts from text. In *Web Intelligence*, pages 1046–1049. IEEE Computer Society, 2006.

[4] E. P. (CLEF-IP). European patents (clef-ip), 2009. [Online; accessed 07-January-2010].

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

[6] W. English. Wikipedia english, 2009. [Online; accessed 12-May-2009].

[7] F. Geraci, M. Pellegrini, M. Maggini, and F. Sebastiani. Cluster generation and labeling for web snippets: A fast, accurate hierarchical solution. *Internet Mathematics*, 3(4):413–443, 2007.

[8] F. P. Giovanna Roda, John Tait and V. Zenz. Clef-ip 2009: Retrieval experiments in the intellectual property domain. In *Working Notes for the CLEF 2009 Workshop*, 2009.

[9] E. J. Glover, D. M. Pennock, S. Lawrence, and R. Krovetz. Inferring hierarchical descriptions. In *CIKM '02: Proceedings of the eleventh international conference on Information and knowledge management*, pages 507–514. ACM Press, 2002.

[10] H. Guan, J. Zhou, and M. Guo. A class-feature-centroid classifier for text categorization. In *WWW '09: Proceedings of the 18th international conference on World wide web*, page 201. ACM Press, 2009.

[11] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK, 2008.

[12] O. D. P. (ODP). Open directory project (odp), 2009. [Online; accessed 29-September-2009].

[13] S. Osinski and D. Weiss. A concept-driven algorithm for clustering search results. *IEEE Intelligent Systems*, 20(3):48–54, 2005.

[14] A. Popescul and L. H. Ungar. Automatic labeling of document clusters, 2000.

[15] D. Radev, H. Jing, M. Stys, and D. Tam. Centroid-based summarization of multiple documents. 2004.

[16] V. Sabol, W. Kienreich, M. Muhr, W. Klieber, and M. Granitzer. Visual knowledge discovery in dynamic enterprise text repositories. In *IV '09: Proceedings of the 2009 13th International Conference Information Visualisation*, pages 361–368. IEEE Computer Society, 2009.

[17] H. Toda and R. Kataoka. A clustering method for news articles retrieval system. In *WWW '05: Special interest tracks and posters of the 14th international conference on World Wide Web*, pages 988–989. ACM Press, 2005.

[18] O. T. C. TREC-9. Ohsumed test collection trec-9, 2000. [Online; accessed 14-December-2009].

[19] P. Treeratpituk and J. Callan. Automatically labeling hierarchical clusters. In *DGO '06: Proceedings of the 2006 International Conference on Digital Government Research*, pages 167–176, 2006.