Tochtermann, Maurer (Eds.): Proceedings of the I-KNOW ’04, Graz
4th International Conference on Knowledge Management
Journal of Universal Computer Science, pp. 353-360, ISSN 0948-6968

Topic Identification: Framework and Application

Benno Stein
(Paderborn University, Germany
stein@upb.de)

Sven Meyer zu Eissen
(Paderborn University, Germany
smze@upb.de)

Abstract: This paper is on topic identification, i. e., the construction of useful labels for sets of documents. Topic identification is essential in connection within categorizing search applications, where several sets of documents are delivered and an expressive description for each category must be constructed on the fly.

The contributions of this paper are threefold. (1) It presents a framework to formally specify the topic identification problem along with its desired properties, (2) it introduces a classification scheme for topic identification algorithms and outlines the respective algorithm of the AIsearch meta search engine, (3) it proposes a hybrid approach to topic identification, which relies on classification knowledge of existing ontologies.

Key words: Topic Identification, Cluster Labeling, Document Categorization, Ontology
Category: H.3 Information Storage and Retrieval

1 Topic Identification in Categorizing Search

Categorizing search means to apply text categorization facilities to retrieval tasks where a large number of documents is returned. Consider for example the use of Internet search engines like Google or Lycos: Given a query they deliver a bulky result list D of documents. Categorizing search means to return D as a set of—a-priori unknown—categories such that thematically similar documents are grouped together.

Categorizing search has attracted much interest recently; its potential has been realized by users and search engine developers in the same way. Its operationalization entails several challenges such as *efficiency* and *nescience*, to mention only two. Efficiency means that category formation must be performed at minimum detention, while nescience means that category formation often happens unsupervised since no predefined categorization scheme is given. At the moment, being in the age before the Semantic Web, clustering technology has achieved considerable success in mastering this ad-hoc category formation task [Stein and Meyer zu Eißen 2002] [see 1].

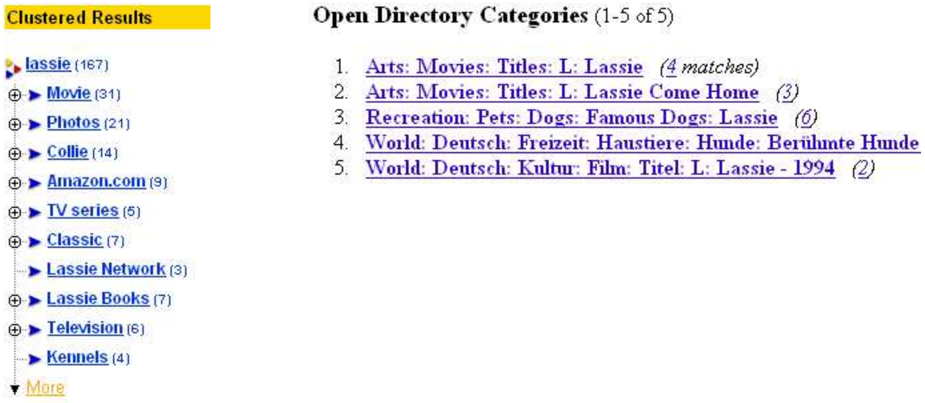
The focus of this paper lies on an aspect of category formation to which less attention has been paid in the past: The identification (construction) of labels that adequately describe the categories that have been found by a clustering algorithm. This problem is called “topic identification” here, but it is also known as “topic finding”, “label identification”, “cluster labeling”, or “category labeling”, and it relates to the problem of keyword extraction as well [Ertöz et al. 2001; Frank et al. 1999; Lagus and Kaski 1999].

[1] The Semantic Web along with its powerful annotation and query concepts based on RDF, RDFS, and DAML+OIL may render the current unsupervised classification efforts superfluous in the medium term. Such a semantically rich approach requires corpora with annotated documents and the access to upper ontologies [Davies et al. 2003; Dill et al. 2003].

At first sight, the adequate labeling of a category seems to be less difficult than its formation. Category formation by clustering is based on a particular document model along with a compatible similarity measure. This naturally extends to cluster labeling, where a cluster C is represented by some kind of a meta document, e. g. the cluster centroid, $\mathbf{c}$, which is built from the document models of all documents in C. Cluster labeling then can be understood as a function $\tau(C)$ that maps onto a set of descriptor terms, which usually is a subset of the index terms in $\mathbf{c}$. Following clustering terminology, this approach should be called "polythetic labeling": The labeling process is based on the simultaneous analysis of several features, the index terms [see 2].

While the polythetic approach is justified for cluster formation, say, when analyzing documents with respect to their similarity [see 3], it is problematic in connection with cluster labeling. Cluster labels are in the role of concept descriptors in a concept hierarchy: Ideally, the cluster labels of a (hierarchical) clustering should represent a conceptualization of the documents in a collection, i. e., they should provide an ontological view. Observe that a concept hierarchy can be understood as the result of a monothetic clustering algorithm: In each clustering step, the most general concept t of the remaining set D' of documents is considered as a feature variable with values $t_1,\ldots,t_k$ according to which the documents in D' are discriminated.

[Figure 1] illustrates the difference between polythetic labeling and the selection of concept descriptors by contrasting the search results of two well-known document retrieval applications: Vivísimo and dmoz. The query term was "Lassie".



Clustered Results

- lassie (167)
- Movie (31)
- Photos (21)
- Collie (14)
- Amazon.com (9)
- TV series (5)
- Classic (7)
- Lassie Network (3)
- Lassie Books (7)
- Television (6)
- Kennels (4)
- More

Open Directory Categories (1-5 of 5)

1. Arts: Movies: Titles: L: Lassie *(4 matches)*
2. Arts: Movies: Titles: L: Lassie Come Home *(3)*
3. Recreation: Pets: Dogs: Famous Dogs: Lassie *(6)*
4. World: Deutsch: Freizeit: Haustiere: Hunde: Berühmte Hunde
5. World: Deutsch: Kultur: Film: Titel: L: Lassie - 1994 *(2)*

Figure 1: *The left-hand side shows the topmost level of the category tree that has been generated by Vivísimo from 150 documents for the query "Lassie"; the category names result from a polythetic labeling algorithm. The right-hand side shows the respective dmoz-categories where Lassie-documents have been found.*

The example of Vivísimo shows that polythetic labeling can produce useful results; however, when unfolding the Vivísimo category tree, several weak points become ob-

[2] This observation was already made in [Sanderson and Croft 1999].

[3] As various analyses have shown, cluster algorithms are able to resemble a categorization scheme found by human editors [Zaho and Karypis 2002; Han et al. 2001].

vious: repeated category names, non-specializing sub-categories, or meaningless category names. By contrast, the dmoz categories are maintained by human editors, and the query results reflect the editors' ontological understanding of the world. The *automatic derivation* of such a genuine hierarchy of categories for a given document collection D is very difficult: researchers have been investigating the combination of sophisticated techniques such as query expansion, local context analysis (LCA), subsumption analysis, collocation analysis, or latent semantic indexing (LSI) [Soto 1999]. The results are often unsatisfactory—not to mention the runtime complexity. Thus, the current practice of categorizing search engines is to realize topic identification in an "ad-hoc-manner", by some kind of index term selection.

The paper in hand contributes right here. It renders the topic identification problem more precisely and introduces the topic identification algorithm of the AIsearch meta search engine. Moreover, the paper proposes a hybrid approach, which relies on the classification knowledge of existing ontologies, and which has the potential to overcome the problems of current topic identification algorithms.

2 A Formal Framework

Let D designate a set of documents. A categorization $\mathcal{C} = \{C \mid C \subseteq D\}$ of D is a division of D into sets for which the following condition holds: $\bigcup_{C_i \in \mathcal{C}} C_i = D$. $\mathcal{C}$ is called an exclusive categorization if $C_i \cap C_{j \neq i} = \emptyset$, $C_i, C_j \in \mathcal{C}$, and a non-exclusive categorization otherwise. The elements in $\mathcal{C}$ are called categories. Moreover, we assume that no categorization scheme is given, which is the normal case for categorizing search engines: They generate $\mathcal{C}$ by means of some clustering approach.

Several clustering algorithms define a hierarchy or can be applied in a recursive manner, this way defining a hierarchy $H_{\mathcal{C}}$ on $\mathcal{C}$. $H_{\mathcal{C}}$ is a tree whose nodes correspond to the categories in $\mathcal{C}$ from which one is marked as root node. Given two categories, $C_i, C_j, C_i \neq C_j$, we write $C_i \succ C_j$ if the corresponding nodes in $H_{\mathcal{C}}$ lie on a common path emanating at the root and if C_i is closer to the root than C_j.

The elements in D are abstractions of the interesting documents; they have been formed according to some document model. In the following we consider a document data structure $d \in D$ as an ordered set of index terms $W_d = \{w_{d_1}, \ldots, w_{d_n}\}$ along with a set of functions, which map from W_d to $\mathbf{R}^+$, such as the term frequency $tf_d(w)$ or the inverse document frequency $idf(w)$.

Let $W = \bigcup_{d \in D} W_d$ be the entire word set underlying $\mathcal{C}$. Then topic identification means the construction of a function τ that assigns to each element $C \in \mathcal{C}$ a set $T_C \subset W$, say, $\tau(C) \mapsto T_C$. τ is called a labeling.

Let $\mathcal{C}$ be a categorization. Then the following are desired properties of a labeling τ:

1. *Unique.* $\forall_{\substack{C_i, C_j \in \mathcal{C} \\ C_i \neq C_j}} : \tau(C_i) \cap \tau(C_j) = \emptyset$

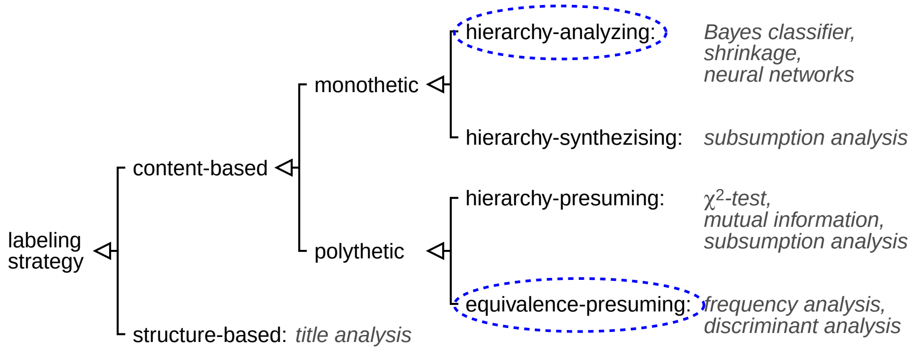
2. *Summarizing.* $\forall_{C \in \mathcal{C}} \, \forall_{d \in C} : \tau(C) \cap W_d \neq \emptyset$

3. *Expressive.* $\forall_{C \in \mathcal{C}}\ \forall_{d \in C}\ \forall_{w \in W_d} : \mathit{tf}_d(w) \leq \mathit{tf}_d(w'),\ w' \in \tau(C)$,
   where $\mathit{tf}_d(w)$ designates the term frequency of some word w in document d.
4. *Discriminating.* $\forall_{\substack{C_i, C_j \in \mathcal{C} \\ C_i \neq C_j}}\ \forall_{w \in W_{C_i}} : \frac{1}{|C_i|}\mathit{tf}_{C_i}(w) \ll \frac{1}{|C_j|}\mathit{tf}_{C_j}(w'),\ w' \in \tau(C_j)$,
   where $W_C = \bigcup_{d \in C} W_d$ designates the word set of category C, and $\mathit{tf}_C(w)$ is the term frequency of w in category C, say, $\mathit{tf}_C(w) = \sum_{d \in C} \mathit{tf}_d(w)$.
5. *Contiguous.* $\forall_{C \in \mathcal{C}}\ \forall_{\substack{w', w'' \in \tau(C) \\ w' \neq w''}}\ \forall_{d \in C}\ \exists_{w_i, w_{i+1} \in W_d} : w_i = w' \wedge w_{i+1} = w''$
6. *Hierarchically Consistent.* $\forall_{\substack{C_i, C_j \in \mathcal{C} \\ C_i \neq C_j}} : C_i \succ C_j \Rightarrow P(w_i|w_j) = 1 \wedge P(w_j|w_i) < 1$,
   where $w_i \in (W_{d_i} \cap \tau(C_i))$, $w_j \in (W_{d_j} \cap \tau(C_j))$, $d_i \in C_i$, $d_j \in C_j$.
7. *Irredundant.* $\forall_{C \in \mathcal{C}}\ \forall_{\substack{w', w'' \in \tau(C) \\ w' \neq w''}}$: w' and w'' are not synonymous.

Remarks. The stated properties formalize ideal constraints which, in the real world, can only be approximated. Note that merely Property 6 requires the existence of a category tree $H_{\mathcal{C}}$. Finally note that hierarchical consistency and irredundancy are practically impossible to achieve if no external knowledge is provided.

3 Topic Identification: Classifying Existing Approaches

There is only little research which directly relates to topic identification. In the previous section we thus provided a framework with desired properties for a labeling τ—now we give a classification scheme for the existing approaches from an operational point of view. This scheme, illustrated in [Figure 2], does also include strategies that were not meant for labeling (topic identification) purposes in first place, but that can be utilized straightforwardly in this respect.



labeling strategy
content-based
structure-based: title analysis
monothetic
polythetic
hierarchy-analyzing:
Bayes classifier, shrinkage, neural networks
hierarchy-synthezising:
subsumption analysis
hierarchy-presuming:
χ2-test, mutual information, subsumption analysis
equivalence-presuming:
frequency analysis, discriminant analysis

Figure 2: *A classification scheme for labeling strategies for document sets; examples for the respective strategy are noted behind the leafs of the tree. The two encircled strategies form candidates that ideally complement each other within a hybrid labeling approach (cf. Section 4).*

At the first level we distinguish between content-based and structure-based labeling strategies. The former rely on the words of the interesting documents, say, what is called a document model in information retrieval. The latter analyze a document with respect to salient text elements, such as a title, an existing keyword list, or typographic emphasis. Content-based and structure-based strategies can be combined, of course. Among others, a structure-based strategy is pursued in the well-known Scatter/Gather system [Cutting et al. 1992] and in systems for automatic text summarization.

At the second level the content-based strategies divide into monothetic and polythetic labeling approaches; their conceptual differences along with the related impacts have been discussed at the outset. Polythetic labeling is used in most document categorization approaches and can hence can be regarded as the prevalent topic identification strategy [Zamir and Etzioni 1998; Zaho and Karypis 2002; Ertöz et al. 2001; Stein and Meyer zu Eißen 2002], to mention only a few. Another important application of polythetic labeling stems from the field of self-organizing maps, whose interpretation is quite difficult. Here, labeling algorithms are used to make the structure and the information available in the map easier to understand [Lagus and Kaski 1999; Rauber 1999]. If no hierarchical order (a hyponymy) is presumed, the labeling construction techniques in polythetic labeling rely on word and document frequencies, quantization errors, and their combination. Otherwise, χ^2-tests, mutual information gain, or subsumption analyses are applied [Popescul and Ungar 2000].

At the third level the monothetic labeling approaches are further differentiated by the underlying knowledge source: hierarchy-analyzing approaches use external knowledge, say, an existing ontology or taxonomy to derive label information. By contrast, hierarchy-synthesizing approaches use the (internal) knowledge of the given document collection; they try to construct concept hierarchies by interrelating categorization and labeling within a single process [Sanderson and Croft 1999]: A label determines a category and vice-versa.

Topic identification that is based on the analysis of an externally provided hierarchy is a new approach for categorizing search engines, which is introduced in [Section 4].

Topic Identification in AIsearch

AIsearch [see 4] is a categorizing meta search engine, which is developed at our institute [Meyer zu Eißen and Stein 2002]. The employed topic identification algorithm could be designated as "weighted centroid covering" and is a polythetic, equivalence-presuming approach. Aside from efficiency its design rationale was a (heuristic) maximization of the properties 1 – 4 mentioned in [Section 2].

As before, let D be a set of documents over a set of words W, let w be a word in W, let $\mathcal{C} = \{C_1, \ldots C_{|\mathcal{C}|}\}$ be a clustering (categorization) of D, and let $\mathit{tf}_C(w)$ denote the term frequency of word w in cluster (category) $C \in \mathcal{C}$. Moreover, let $\kappa : W \times \{1, \ldots, |\mathcal{C}|\} \to \mathcal{C}$ be a function with $\kappa(w, i) = C$ iff C is the cluster with the ith frequent occurrence of word w. For example, $\kappa(w, 1)$ and $\kappa(w, |\mathcal{C}|)$ denote those clusters wherein w occurs most frequently and least frequently, respectively.

The algorithm consists of two major parts. At first, a vector $\mathcal{T}$ of $k \cdot |W|$ tuples $\langle w,\ \mathit{tf}_{\kappa(w,i)}(w)\rangle$, $i \in \{1, \ldots, k\}$, is constructed for the k most frequent occurrences of

[4] www.aisearch.de

a word w in $\mathcal{C}$; $\mathcal{T}$ is sorted in descending order with respect to the term frequencies. Secondly, l different words are assigned to each of the clusters, where in each pass exactly one word is assigned to every cluster (Round-Robin). Observe that by processing the tuples in $\mathcal{T}$ in a top-down manner, the most frequent words in the weighted centroids of the clusters $C \in \mathcal{C}$ are selected (covered)—which advised us to give the algorithm its name.

Algorithm: Weighted Centroid Covering (WCC)
Input: A clustering $\mathcal{C}$.
l. Specifies how many terms make up a label.
k. Specifies how often the same word may occur in the label of different clusters.
Output: A labeling function τ.

```
WCC($\mathcal{C}, l, k$)
 1. $\mathcal{T} = \emptyset$;
    FOREACH $C$ IN $\mathcal{C}$ DO $\tau(C) = \emptyset$;
 2. FOREACH $w$ IN $W$ DO
      FOR $i = 1$ TO $k$
        compute $C = \kappa(w, i)$ from $\mathcal{C}$;
        add tuple $\langle w,\ tf_C(w)\rangle$ to $\mathcal{T}$;
      ENDFOR
    ENDDO
 3. SORT $\mathcal{T}$ descending with respect to the term frequencies;
 4. FOR $labelcount = 1$ TO $l$
      $assigned = 0$; $j = 1$;
      WHILE $assigned < |\mathcal{C}|$ AND $j \leq |\mathcal{T}|$
        let $t_j = \langle w,\ tf_C(w)\rangle$ denote the $j$th tuple of $\mathcal{T}$;
        IF $|\tau(C)| < labelcount$ THEN
          $\tau(C) = \tau(C) \cup \{w\}$;
          delete $t_j$ from $\mathcal{T}$;
          $assigned = assigned + 1$;
        ENDIF
        $j = j + 1$;
      ENDWHILE
    ENDFOR
 5. RETURN $\tau$;
```

The labeling τ generated by WCC fulfills the properties 1 and 2 by definition of κ; the cluster-wise Round-Robin-strategy aims at fulfilling Property 3. A small k as parameter helps fulfilling Property 4. Computing the κ-values (including sorting) is in $O(k \cdot |W| \cdot \log(k \cdot |W|))$; assigning the labels is in $O(l \cdot k \cdot |W|)$. Since k and l are typically bounded by a small constant, the overall complexity is in $O(|W| \cdot \log(|W|))$.

4 Hybrid Topic Identification Using Upper Ontologies

The preparation of search results D in the form of a category tree $H_{\mathcal{C}}$ will be of maximum value for each human information miner, if—but only if(!)—both the related labeling provides an ontological view onto D and the documents in the categories obey the conceptualization of this ontology.

Clearly, even for a large set of search results a category tree $H_{\mathcal{C}}$ can be generated easily by a hierarchical clustering approach [El-Hamdouchi and Willett 1989]. However, the resulting labeling τ, which is based on $H_{\mathcal{C}}$, is usually far away from being a useful taxonomy. In particular, the properties 6 and 7 of the framework cannot be fulfilled.

We believe that the weaknesses of topic identification algorithms in categorizing search engines could be overcome if external classification knowledge were brought in. We now outline the ideas of such an approach where both topic descriptors and hierarchy information from an upper ontology are utilized. Let $\mathcal{C} = C_1, \ldots, C_m$ be the categorization constructed by a clustering algorithm for a set D of search results. Moreover, let $\mathcal{O} = O_1, \ldots, O_l$ be a reference categorization of a set of documents $D_{\mathcal{O}}$, let $\tau_{\mathcal{O}}$ be a labeling of $\mathcal{O}$ that provides an ontological view onto $D_{\mathcal{O}}$, and let the documents in the O_j obey the conceptualization of this ontology. Then, topic identification is based on the following paradigms:

1. Initially, no hierarchy (refines-relation) is presumed among the $C \in \mathcal{C}$. This is in accordance with the observations made in [Ertöz et al. 2001].

2. Each category $C \in \mathcal{C}$ is associated to its most similar set $O \in \mathcal{O}$. If the association is unique, $\tau_{\mathcal{O}}(O)$ is selected as category label for C.

3. Categories which cannot be associated uniquely within $\mathcal{O}$ are treated by a polythetic, equivalence-presuming labeling strategy in a standard way.

In essence, finding a labeling for a categorization $\mathcal{C}$ using an ontology $\mathcal{O}$ means to construct a hierarchical classifier, since one has to map the centroid vectors of the clusters $C \in \mathcal{C}$ onto the best-matching $O \in \mathcal{O}$. Note that a variety of machine learning techniques has successfully been applied to this problem; they include Bayesian classifiers, SVMs, decision trees, neural networks, regression techniques, and nearest neighbor classifiers.

5 Summary and Current Work

Topic identification means to construct a list with few words (= label) to characterize a set of documents. In categorizing search we have a set $\mathcal{C}$ of such document sets, and the construction of a set of meaningful labels (= labeling τ) is both important and difficult, since label construction may not happen isolated for each category $C \in \mathcal{C}$. The paper in hand presents a framework with desired properties of such a labeling τ. Moreover, it classifies existing approaches that may be used for topic identification purposes, it gives an algorithmic specification of the topic identification algorithm in AIsearch, and it proposes a topic identification strategy that relies on existing ontologies.

The new approach has the potential to overcome the problems of current topic identification algorithms—in particular, it solves the problem of term selection and hierarchy construction—and our experiments are promising in this respect. Nonetheless, we have not presented a comparative analysis here for the following reasons: (1) Our experiments relate to selected parts of the dmoz-ontology, and there is the question to what extent they may be generalized; (2) there is no simple measure to compare topic identification algorithms. A comprehensive analysis requires a detailed description of the underlying queries, document collections, and linguistic measures, which goes beyond the scope of this paper[see 5].

Of course, the new approach also rises new questions; perhaps the most crucial is the question of how to obtain a useful and "complete" upper ontology.

References

[Cutting et al. 1992] Douglass R. Cutting, David R. Karger, Jan O. Pedersen, and John W. Tukey. Scatter/Gather: A cluster-based approach to browsing large document collections. In *Proceedings of the 15th ACM SIGIR Conference*, pages 318–329. ACM Press, 1992.

[Davies et al. 2003] John Davies, Dieter Fensel, and Frank van Harmelen. *Towards the Semantic Web: Ontology-Driven Knowledge Management*. John Wiley & Sons, New York, 2003.

[Dill et al. 2003] S. Dill, N. Eiron, D. Gibson, D. Gruhl, R. Guha, A. Jhingran, T. Kanungo, S. Rajagopalan, A. Tomkins, J. A. Tomlin, and J. Y. Zien. SemTag and Seeker: Bootstrapping the Semantic Web via Automated Semantic Annotation. In *Procc of the 12th Int. Conference on the WWW*, pages 178–186. ACM Press, May 2003.

[El-Hamdouchi and Willett 1989] A. El-Hamdouchi and P. Willett. Comparison of Hierarchic Agglomerative Clustering Methods for Document Retrieval. *The Computer Journal*, 32(3): 220–227, 1989.

[Ertöz et al. 2001] L. Ertöz, M. Steinbach, and V. Kumar. Finding Topics in Collections of Documents: A Shared Nearest Neighbor Approach. In *Proceedings of the Text Mine '01, Workshop on Data Mining, 1st SIAM International Conference on Data Mining*, 2001.

[Frank et al. 1999] E. Frank, G. W. Paynter, I. H. Witten, C. Gutwin, and C G. Nevill-Manning. Domain-Specific Keyphrase Extraction. In *Proc. of the 16th International Joint Conference on Artificial Intelligence*, pages 668–673. 1999.

[Han et al (2001] Eui-Hong Han, George Karypis, and Vipin Kumar. Text Categorization Using Weight Adjusted k-Nearest Neighbor Classification. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 53–65, USA, 2001. University of Minnesota.

[He et al. 2001] X. He, C. H. Q. Ding, H. Zha, and H. D. Simon. Automatic Topic Identification Using Webpage Clustering. In *Proc. of ICDM'01*, pages 195–202, November 2001.

[Lagus and Kaski 1999] K. Lagus and S. Kaski. Keyword Selection Method for Characterizing Text Document Maps. In *Proc. of ICANN'99*, pages 371–376. IEEE, 1999.

[Meyer zu Eißen and Stein 2002] S. Meyer zu Eißen and B. Stein. The AISEARCH Meta Search Engine Prototype. *Proc. of the 12th Workshop on Information Technology and Systems (WITS 02), Barcelona, Spain*. December 2002.

[Popescul and Ungar 2000] A. Popescul and L. H. Ungar. Automatic Labeling of Document Clusters. `http://citeseer.nj.nec.com/popescul00automatic.html`, 2000.

[Rauber 1999] A. Rauber. LabelSOM: On the Labeling of Self-Organizing Maps. In *Proc. of IJCNN'99*. IEEE, July 1999.

[Sanderson and Croft 1999] M. Sanderson and W. B. Croft. Deriving Concept Hierarchies from Text. In *Research and Development in Information Retrieval*, pages 206–213, Berkley, 1999.

[Soto 1999] R. Soto. Learning and performing by exploration: label quality measured by latent semantic analysis. In *Proceedings of CHI'99*, pages 418–425. ACM Press, 1999.

[Stein and Meyer zu Eißen 2002] B. Stein and S. Meyer zu Eißen. Document Categorization with MAJORCLUST. *Proc. of the 12th Workshop on Information Technology and Systems (WITS 02), Barcelona, Spain*, pages 91–96. December 2002.

[Zaho and Karypis 2002] Y. Zaho and G. Karypis. Criterion Functions for Document Clustering: Experiments and Analysis. Tech. Report 01-40, Univercity of Minnesota, 2002.

[Zamir and Etzioni 1998] O. Zamir and O. Etzioni. Web Document Clustering: A Feasibility Demonstration. In *Proc. of the 21st ACM SIGIR conference*, pages 46–54, USA, 1998.

[5] In an upcoming technical paper we will report on a comparison of the topic identification algorithms of different categorizing search engines.