Handbook of Research on Text and Web Mining Technologies

Min Song
New Jersey Institute of Technology, USA

Yi-fang Brook Wu
New Jersey Institute of Technology, USA

Volume I

Information Science REFERENCE

INFORMATION SCIENCE REFERENCE
Hershey · New York

Director of Editorial Content:	Kristin Klinger
Director of Production:	Jennifer Neidig
Managing Editor:	Jamie Snavely
Assistant Managing Editor:	Carole Coulson
Typesetter:	Chris Hrobak
Cover Design:	Lisa Tosheff
Printed at:	Yurchak Printing Inc.

Published in the United States of America by
Information Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue, Suite 200
Hershey PA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: http://www.igi-global.com

and in the United Kingdom by
Information Science Reference (an imprint of IGI Global)
3 Henrietta Street
Covent Garden
London WC2E 8LU
Tel: 44 20 7240 0856
Fax: 44 20 7379 0609
Web site: http://www.eurospanbookstore.com

Library of Congress Cataloging-in-Publication Data

Handbook of research on text and web mining techologies / Min Song and Yi-Fang Wu, editors.

p. cm.

Includes bibliographical references and index.

Summary: "This handbook presents recent advances and surveys of applications in text and web mining of interests to researchers and end-users "--Provided by publisher.

ISBN 978-1-59904-990-8 (hardcover) -- ISBN 978-1-59904-991-5 (ebook)

1. Data mining--Handbooks, manuals, etc. 2. Web databases--Handbooks, manuals, etc. I. Song, Min, 1969- II. Wu, Yi-Fang, 1970-

QA76.9.D343H43 2008

005.75'9--dc22

2008013118

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.

All work contributed to this book set is original material. The views expressed in this book are those of the authors, but not necessarily of the publisher.

Chapter XX
Thesaurus-Based Automatic Indexing

Luis M. de Campos[1]
University of Granada, Spain

Juan M. Fernández-Luna
University of Granada, Spain

Juan F. Huete
University of Granada, Spain

Alfonso E. Romero
University of Granada, Spain

ABSTRACT

In this chapter, we present a thesaurus application in the field of text mining and more specifically automatic indexing on the set of descriptors defined by a thesaurus. We begin by presenting various definitions and a mathematical thesaurus model, and also describe various examples of real world thesauri which are used in official institutions. We then explore the problem of thesaurus-based automatic indexing by describing its difficulties and distinguishing features and reviewing previous work in this area. Finally, we propose various lines of future research.

INTRODUCTION

Automated text categorization (Sebastiani, 2002) is a successful subfield of information management and has to date been the subject of more than six hundred research publications during more than forty-five years of work (Sebastiani & Gabrilovich, 2007). Although it is intrinsically a research field (many scientists and engineers attempt to build good *text classifiers* using different methods) as part of text mining, this discipline can also be helpful for obtaining knowledge such as document summarization, concept or entity extraction, sentiment analysis, or document clustering from documents. Even in the field of information retrieval (van Rijsbergen, 1979) it is interesting for documents to be previously categorized into a list of classes so that they may be retrieved more accurately using associated keywords as meta-information.

Text documents and natural language both share the common problems of lexical ambiguity or polysemy (i.e. words or expressions having more than one meaning and which is a special case of homonymy) and synonymy (different terms or expressions for the same concept). Additionally, the most common representation used for a text document is the bag of words approach (a model similar to "first-order word approximation" used by Shannon in 1948) and this reduces a document to a list of unrelated terms usually resulting in loss of contextual meaning and structure of certain expressions present in the text.

One tool which has been intrinsically designed to avoid ambiguities of any class is a thesaurus. This is a set of terms with orthogonal meanings and a set of the hierarchical relationships between them. A thesaurus can be very useful in different areas of text mining (Berry, 2003) by removing ambiguity and identifying a document's context. In this chapter, we present some thesaurus basics, a formal characterization, and several examples of real world thesauri. We will then focus on the problem of automatic indexing on a domain of categories defined on a thesaurus.

THESAURI BASICS

Definitions

Broadly speaking, a *thesaurus* consists of a set of terms (which are relevant to a certain domain of knowledge) and a set of relationships between them. Its main aim is to represent concepts without ambiguity in order to avoid confusion and misunderstanding.

The basic unit of a thesaurus is often called the *descriptor*[2]. A descriptor is a word or phrase which identifies an important notion in that domain of knowledge, i.e. it designates essential entities in the area covered by the thesaurus.

Other basic thesaurus units are the *non-descriptors*. These are words or expressions which basically denote the same notion as a descriptor in the thesaurus language.

A thesaurus also includes the following three types of semantic relationships:

- *Hierarchical relationships* involving one descriptor and a more specific or broader one.
- *Equivalence relationships* between non-descriptors and descriptors, listing the equivalent terms for a certain concept, and the possible uses of a descriptor. The equivalence relationships may in fact cover relationships of various types: identical, similar or opposite meanings, and even inclusion.
- *Associative relationships* between descriptors, which can also be of various kinds: cause and effect, agency or instrument, concomitance, constituent elements, location, etc. They generally link two descriptors that do not meet the criteria for either equivalence or hierarchical relationships, and are used to suggest another descriptor that would be helpful for the thesaurus user to search with.

In the first case, the hierarchy defined by the thesaurus specifies the BT (*broader term*) and NT (*narrower term*) relationships. The NT designates a more specialized descriptor for a particular one. For each BT, there is the corresponding NT relationship (they are reciprocal). In other words, if the broader term of "A" is "B", then the narrower term of "B" is "A". If a descriptor has no broader term, it is sometimes called a *top term*. On the other hand, if a descriptor has no narrower term, it is often called a *basic descriptor*.

In the second case, the equivalence relationships are UF (*used for*) and USE: UF between the descriptor and the non-descriptor(s) it represents, and USE between a non-descriptor and the descriptor which replaces it.

For the third case, a thesaurus specifies a *related term* relationship (RT). This is a symmetrical relationship: if a descriptor "A" is related to "B" by means of an RT relation, then "B" and "A" should also be in RT.

Certain thesauri include a *scope note* (SN) for several descriptors which defines or limits the specific use of a term. Following Steinberger, Pouliquen and Ignat's distinction (2003), we will divide thesauri into two different kinds: conceptual thesauri, with abstract and conceptual terms as descriptors (e.g. EUROVOC), and natural language thesauri, which tend to be more specific covering a certain area of knowledge (e.g. AGROVOC or MeSH).

Thesaurus Formalization

A thesaurus may be formalized as an eight-tuple

$$(\Omega, \Delta, \Gamma, W_{\Delta}, W_{\Gamma}, USE, RT, BT),$$

where the sets $\Omega = \{\omega_1,..., \omega_n\}$, $\Delta = \{\delta_1,..., \delta_m\}$ and $\Gamma = \{\gamma_1, ..., \gamma_k\}$ represent the terms (words), descriptors and non-descriptors in the thesaurus, respectively.

There is a map $W_{\Delta} : \Delta \rightarrow 2^{\Omega} \setminus \{\emptyset\}$, where 2^{Ω} is the set of all subsets of Ω. Similarly, there is another map $W_{\Gamma} : \Gamma \rightarrow 2^{\Omega} \setminus \{\emptyset\}$. Clearly, $W_{\Delta}(\delta)$ (resp. $W_{\Gamma}(\gamma)$) denotes the set of terms associated to a descriptor δ (respectively a non-descriptor γ).

There is also a map $USE : \Gamma \rightarrow \Delta$, such that $\forall\ \gamma \in \Gamma,\ USE(\gamma) \in \Delta$ is the descriptor associated with the non-descriptor γ. Therefore, the inverse map UF can also be defined as $UF(\delta) = USE^{-1}(\delta)$ (the set of non-descriptors associated with the descriptor δ).

The relation map $RT : \Delta \rightarrow \Delta$ is defined between pairs of descriptors, verifying $\forall\ \delta_1, \delta_2 \in \Delta$, if $RT(\delta_1) = \delta_2 \Rightarrow RT(\delta_2) = \delta_1$. Thus, the binary relation corresponding to RT is symmetrical.

Finally, there is another function $BT : \Delta \rightarrow \Delta \cup \{\emptyset\}$ (where $\emptyset$ represents the *empty descriptor*) such that $BT(\delta) \in \Delta$ is the broader descriptor containing δ and $BT(\delta) = \emptyset$ means that the descriptor δ is not contained in a more general one (i.e. it is a top term). $NT(\delta) = BT^{-1}(\delta)$ is the set of narrower descriptors which are contained in δ. If $BT^{-1}(\delta) = \emptyset$, then the descriptor δ does not contain a more specific descriptor (it is a basic descriptor). If $BT^{-1}(\delta) \neq \emptyset$, we say that δ is a non-basic or complex descriptor. More generally, when a descriptor δ is polyhierarchical (as in EUROVOC, for instance), then $BT(\delta)$ is not a single descriptor but a subset of descriptors, hence BT is not a function but a correspondence.

The function/correspondence BT must satisfy the property

$$BT(\delta)\ o...k...o\ BT(\delta) \neq \delta,\ \forall\ \delta \in \Delta, \forall\ k=1,2,...,$$

where *o* represents composition. This property guarantees that the *BT* relationships constitute a true hierarchy.

Defining Thesauri

Most of the thesauri mentioned in this chapter are public and freely available for non-commercial purposes, and most can be obtained in several non-standard formats: XML, ASCII (plain text), PDF, etc. Defining a thesaurus following a standard language is very important if certain applications are needed (navigating and consulting the thesaurus, and in order to easily obtain descriptors and relationships).

The Simple Knowledge Organisation System or SKOS (Miles and Brickley, 2005), developed by the W3C, is a family of formal languages for thesaurus representation, structured controlled vocabulary and any other taxonomy. SKOS is built on RDF and RDFS, and its main objective is to enable easy publication of controlled structured vocabularies for the Semantic Web. SKOS is currently under development but it is expected to be a proposed recommendation for the W3C in the first quarter of 2008.

REAL WORLD THESAURI

There are several industrial and socio-political thesauri which are used in different domains of knowledge. Some of these are presented in the following section with a description of their distinguishing features.

EUROVOC

EUROVOC (Office for Official Publications of the European Communities, 1995) is an official thesaurus of the European Union. It covers the fields in which the European Communities are active, with its main aim being to provide a means of indexing the official documents in the documentation systems of European institutions and their users. It is used by the European Parliament, the Office for Official Publications of the European Communities, national and regional European parliaments, and other non-EU countries.

Since EUROVOC is a plurilingual thesaurus, different editions of it have been published in the twenty-one official EU languages. All of them share the same descriptors, hierarchical relationships and associative relationships, making it easy to provide “conceptual indexation” and to search for a document, regardless of the language in which it is written.

The first level of the hierarchy is called the thematic field – twenty-one in EUROVOC. The second level is the microthesaurus (containing 127 microthesauri), and each thematic field comprises certain microthesauri.

Its current version (Version 4.2) consists of 6645 descriptors, with 519 top terms, 6669 hierarchical relationships (BT/NT), and 3636 associative relationships (RT). See Table 1 for a comparison of thesauri sizes.

As mentioned previously, some of the descriptors are polyhierarchical (they can have more than one descriptor as a broader term), and RT relationships are incompatible with hierarchical ones. EUROVOC is a conceptual thesaurus in the sense that very different fields in socio-political life (geographical areas, medical terminology, agriculture, etc.) can be indexed with it.

AGROVOC

AGROVOC (FAO World Agricultural Information Center, 1998) is a multilingual, structured and controlled vocabulary designed to cover the terminology of all subject fields in agriculture, forestry, fisheries, food and related domains (e.g. environment). It has been developed since 1980 by the Food and Agriculture Organization of the United Nations (FAO), and has been translated into 17 languages. It is free for non-commercial use. It is similar in structure to EUROVOC and uses a new relationship called "spatially included in" (for geographical descriptors). The English edition of AGROVOC contains 28705 descriptors and 10927 non-descriptors, and 32176 BT/NT and 27589 RT relationships. AGROVOC is used all over the world in about ninety countries, mostly for indexing and retrieving data in agricultural information systems.

NAL Thesaurus

The NAL Agricultural Thesaurus (NALT) (National Agriculture Library, 2002) was built in 2002 by the National Agricultural Library to meet the needs of the USDA, Agricultural Research Service (ARS). The subject scope of agriculture is broadly defined in the NALT and includes terminology in the supporting biological, physical and social sciences. Biological nomenclature comprises a majority of the terms in the thesaurus and is located in the "Taxonomic Classification of Organisms" Subject Category. Political geography is mainly described at the country level.

This thesaurus is currently used on several sites. In 2003, NAL implemented the thesaurus as the controlled indexing vocabulary for their in-house bibliographic database, AGRICOLA, in conjunction with the implementation of a new electronic library management system. In 2004, E-Extension, the prototype database of state cooperative extension publications, chose the NAL Agricultural Thesaurus as their controlled vocabulary. Other non-USDA sites use the thesaurus, such as the Agricultural Network Information Center (AgNIC), which adopted it in 2002. AgNIC partners use the thesaurus as a vocabulary for indexing and as a search aid to their Web sites through Web services on their portal.

It is organized into 17 subject categories, indicated by the "Subject Category" designation in the thesaurus, so that the thesaurus can be browsed in a specific discipline or subject area. It includes 42326 descriptors, 26238 non-descriptors, 44545 BT/NT hierarchical relationships, and 17324 symmetrical RT relationships.

Table 1. Comparative numbers for real thesauri: the number of descriptors and non-descriptors (only English ones if the thesaurus is multilingual) are shown, together with the number of relationships (the size of the graph represented by the thesaurus).

Name	EUROVOC	AGROVOC	NAL
Descriptors	6645	28705	42326
Relationships (BT/NT+RT)	10305	59765	61869
Non-descriptors	6769	10927	26238

The NAL Agricultural Thesaurus is also freely available online from the NAL website and as a Web service to other Web-connected programs.

MeSH

MeSH (National Library of Medicine, 1986) is a significant thesaurus in biomedicine and was built by the American National Library of Medicine (NLM) in 1960. It has subsequently been updated and translated into many different languages. It has three main aims: indexing biomedical literature (published in Index Medicus), cataloguing monographs and multimedia content stored in the NLM, and presenting a vocabulary for the Index Medicus user, in order to make precise and accurate searches in the index.

MeSH contains over 24,767 headings (descriptors in MeSH terminology) and the three types of previously described relationships: hierarchical or BT/NT (MeSH is divided into 16 trees), synonymous (over 97,000 non-descriptors, which are called entry terms here), and related terms. MeSH also includes several additional features which distinguishes it from other thesauri:

- Qualifiers (also known as subheadings) used for indexing and cataloging in conjunction with descriptors
- *Publication types:* type of publication or type of study, they indicate what the indexed item is, rather than what it is about
- *S*everal added semantic relationships (*concept*, *consider also*, *entry combination*, etc).

An introduction to the MeSH thesaurus and to biomedical information in general can be found in Hersh (2003).

THESAURUS-BASED AUTOMATIC INDEXING

Text Categorization

The task of automatic indexing from a thesaurus-based set of categories can be defined as a problem of text document categorization in a set of hierarchical classes. Since Maron (1961), a lot of work has been devoted to solving the problem of text categorization in general (see Sebastiani, 2002, or Sebastiani & Gabrilovich, 2007).

In a formal way, given a set of documents $\mathcal{D} = \{d_1,..., d_{|\mathcal{D}|}\}$ and a set of categories $C = \{c_1, ..., c_{|C|}\}$, the main goal of text categorization is to build a classifier, i.e. a map $f\colon \mathcal{D} \rightarrow C$, in order to assign a category to each document. This is called *single-label categorization*. Many approaches to this problem have performed extremely well: probabilistic methods, neural networks, decision trees or support vector machines (see Dumais et al., 1998, for good examples).

A more complex case is building a *multi-label* classifier, where several categories can be assigned to a document, i.e. a map $f : \mathcal{D} \rightarrow 2^C$. This problem can be reduced (Sebastiani, 2002) to the previous one by building a set of $|C|$ binary classifiers $f_i\colon D \rightarrow \{c_i, \neg c_i\}$ $(i = 1,...,|C|)$. It should be observed that there are no restrictions regarding the number of labels assigned to a certain document, although in practice it is quite common to limit this number.

Another approach for performing multi-label categorization is *ranking categories*. In multi-label *hard* categorization, a set of labels is assigned for each document (the labels which do not appear in the result are supposed to be non-important to the document). A simpler approach is to return the entire set C for each document, ranked in order of decreasing value of appropriateness to the document. We will call this appropriateness value *CSV* which stands for *Categorization Status Value* (Sebastiani, 2002). A document being classified in such a system will therefore return the following set of pairs:

$\{(c_i, CSV_i) \,/\, i=1,..,|C|\}$

In order to present the categorization results to the user, it is more usual to use this method and it is better to implement if a semi-automated indexing system is being built. In this approach, the ranked set of categories would be presented to the human indexer, sorted by CSV, and they would then decide the appropriate categories for the current document being classified.

On the other hand, building a completely automated multi-label classifier system from a ranking category system is easy by considering a threshold τ_i for each category. Having obtained the list of ranked categories for a certain document being classified, the category c_i is assigned to a document if $CSV_i \geq \tau_i$.

Text Categorization Approaches

There are three different approaches to text categorization:

1. **Machine-learning approach.** In this approach, we previously have a set of pre-classified documents. This dataset is then split into a *training dataset* and a *validation dataset*, a classifier is learnt using the training examples, and its accuracy is checked using the validation set. This is called the classic "train-test" procedure, although more advanced methods can be used instead (cross-validation, leave-one-out, etc.). The important point is that *we use pre-classified examples to build the classifier*. This is of course a case of *supervised learning*.
2. **Clustering approach.** Here no training examples are given, and we then build groups (clusters) of documents based only on their similarity. For this approach, we only need to define a clustering approach that generally relies on a definition of distance between documents. In this case, we talk about *unsupervised learning*.
3. **Mixed approach,** where both previous methods are combined to obtain better results.

We will now place our problem in the field of text categorization.

Thesaurus Indexing as a Text Categorization Problem

We can define the problem of automatic indexing in a thesaurus-based set of categories as *a problem of text categorization on the set of classes defined by a thesaurus* (the set of descriptors). Because documents are usually assigned several descriptors, we are therefore dealing with a multi-label categorization problem.

Considered within a pure machine learning approach, and starting from a training set of classified documents, solving this problem leads us to the following solutions:

- **Building a flat multi-label classifier.** Here the categories (descriptors) are supposed to be independent and our problem is treated as a classical text categorization approach. This solution presents the following drawbacks:
    - The training set *will not be complete* (i.e. there will not be at least one positive example for each category), and consequently a category with no positive examples will never be returned in the results.
    - No influence between descriptors is considered, although this knowledge would probably be used by a human indexer.
- **Building a hierarchical multi-label classifier**. Some approaches have been published which attempt to model a set of categories with a predefined hierarchy (see Dumais and Chen, 2000). Use of class relationships can partially prevent the two previous drawbacks but will not generally remove them. If the size of the thesaurus is large enough, some parts of the tree representing the hierarchy will be scarcely populated or even empty, leading us again to the first previous problem. On the other hand, the graph represented by a thesaurus is not necessarily a tree or a forest (which are the kinds of graphs considered in hierarchical classifiers), and then only the BT/NT relationships will be taken into account, excluding the RT relationships.

It should be noted that neither of the two previous approaches uses the non-descriptors, or the associative relationships, or the text contained in the descriptors and non-descriptors. On the other hand, thesauri have proved to be a great source of information, and it has been shown in de Campos, Fernandez-Luna, Huete and Romero (2007) that a classifier using only textual information and the relationships from the thesaurus can achieve reasonable results in a pure, unsupervised approach.

The most appropriate way to solve this problem is to combine a machine learning approach with an unsupervised (clustering) approach which uses the textual information from the thesaurus. In line with Golub (2006), we will say that this solution is a *mixed approach*. In our opinion, there are three possible realizations of a mixed approach:

- *Building a (hierarchical) classifier using the textual information of descriptors and non-descriptors as additional training documents.* This solves the problem of lack of information in some of the categories, and seems to be the easiest way of adapting this problem to a machine learning approach, only requiring a methodology to represent the textual information of the thesaurus in the form of training data.
- *Boosting an unsupervised classifier with trained data.* In the case of absence of training information for a given category, we will use the data from the thesaurus to assign a document to that category. With more pre-classified documents assigned to a descriptor, the information from the unsupervised classifier will become less relevant (or will even be ignored) and only supervised information will be considered. This scheme is used for classification in Web directories in Adami, Avesani and Sona (2005).
- *Building an ad-hoc classifier for thesauri.* Neither of the two previous solutions are *pure* mixed approaches since they are considered to be an adaptation of existing methods. One can of course build an *ad-hoc* model of a manual indexer over a thesaurus, getting rid of all the previously considered problems.

Problem Difficulties

This peculiar problem differs from the classical approach of text categorization by presenting the following difficulties:

- *High dimensionality* of the set C of categories: as we have seen in the previous section, we are managing several thousand categories. The standard test collections for text categorization (where the existing approaches are tested) have a small number of categories (none of the classical test collections have more than a thousand categories, with most having fewer than a hundred).
- *The problem of unbalanced data:* the number of training documents associated to each class may be very different, and this will cause problems for any supervised learning algorithm.
- *The lack of specific test collections*: to date, there has been no freely available test corpus of documents classified over a thesaurus (with the possible exception of OHSUMED; see Hersh et al., 1994). This makes it extremely difficult to test new approaches because of the absence of a *standard* comparison procedure.
- *The problem of feature selection:* a thesaurus itself contains a great deal of information. Which is useful for making more accurate classifications? Should certain kinds of relationships not be considered? Should non-descriptor terms be considered as a source of meta-information for a certain category? In addition, standard feature selection methods for documents might not be useful here because of the hierarchical relationships in the set of classes (two classes can be disjoint in the sense that they are different, but semantically related by an "is a" or other semantic relationship). In Moskovitch et al. (2006) a method called *hierarchical mutual information* is proposed for this purpose.
- *The problem of comparison:* when discussing a new categorization method, a baseline is a very simple approach for the problem which is being solved, something which is supposedly worse than our proposal. Classical methods such as Naïve Bayes (Maron, 1961) are sometimes used for this purpose. At other times, a well-working model is used as this baseline. What baseline is suitable for this kind of experimentation? We will discuss a simple solution at a later stage.
- *The problem of evaluation:* this is a very important point for making conclusive results. If the problem is considered to be a category-ranking problem, a very common method of evaluation is based on the precision/recall measures, such as the MAP measure (average precision at the standard eleven points of recall) or the F-measure (van Rijsbergen, 1979). This type of evaluation procedure does not take into account the semantic links between categories. For example, let us imagine that a document in the test set is classified to a category A (which is the broader term of B). If A does not appear in the results of the correct classes of this document but B does, using pure precision/recall evaluation, category A will be considered as wrong even when it is almost the desired category with a certain degree of generality. There is no general criterion for solving this problem. For example, in Moskovitch et al. (2006) evaluation is carried out using their own proposed measures (hierarchical micro recall and hierarchical micro precision). A more in-depth study of the different alternatives for the evaluation of this problem can be found in Sun, Lim and Wee-Keong (2003).
- *The problem of related class influence:* as in hierarchical classification systems, a class can be associated to a certain document by means of only hierarchy-related classes. For example, if all the narrower terms of a descriptor seem to be relevant for a certain document, the system should

instead decide to return the broader descriptor. This means that the training set does not need to be complete (in the sense of containing at least one document of each class), and therefore, a certain descriptor can be returned in the results, even if it does not appear in the entire training set. We have previously mentioned a partial solution to this problem which uses a hierarchical classifier.

Related Work in Automated Indexing

Various references for published work on automatic indexing in a thesaurus domain are given:

- In Montejo-Ráez (2005), an indexer is built for the DESY (DESY Library, 1996) thesaurus (a thesaurus in the high energy physics domain used at the CERN) using several classification methods. A shorter form of this work is also presented in Montejo-Ráez (2002).
- An old approach to indexing in EUROVOC (without considering the structure of the thesaurus) can be found in Marjorie & Hainebach (1996).
- In Steinberger (2001) the automatic indexer for EUROVOC used in the European Parliament is described and evaluated. In Steinberger (2000) and Steinberger et al. (2003), the authors propose an algorithm for automatic indexing of EUROVOC documents in a multilingual environment. This automatic indexing method is also used as a tool to calculate multilingual document similarity in Steinberger, Pouliquen and Hagman (2002).
- A Bayesian network-based approach where no previous knowledge is given to the system except the thesaurus itself is proposed in de Campos et al. (2007) to index parliamentary initiatives in EUROVOC.
- In Medelyan (2005), an algorithm to assign keywords from a thesaurus is presented and tested in AGROVOC. It is also described in Medelyan & Witten (2005; 2006).

A Simple Baseline: Vector Space Model

The Vector Space Model (VSM) is one of the most successful models in information retrieval. It was first proposed by Salton, Wong and Yang (1975) in the 1970s and is still used today. Each document is represented as an *n*-tuple of non-negative real numbers. Each coordinate corresponds to one of the *n* different terms present in the collection. The value of the coordinate describes the importance of the term in the document, i.e. a very important term in the document has a higher coordinate value than less important ones. Let D be the number of documents in the collection, and D_i the number of documents where the term i appears. The value (weight) of the importance of the term i in the document j (the i-th component of the j-th document vector) is denoted by w_{ij}, and is often defined as:

$$w_{ij} = tf_{ij} idf_i = \frac{f_{ij}}{\max_k f_{kj}} \log \frac{D}{D_i},$$

where f_{ij} is the absolute frequency of the i-th term in the j-th document. On the other hand, idf_i stands for inverse document frequency of i-th term, the logarithm of the number of documents divided by the number of documents where this term appears, and measures the rarity of the term in the collection. The similarity between two documents is computed as the cosine of the angle between the two vectors:

$$sim(d_i, d_j) = \frac{\langle d_i, d_j \rangle}{\|d_i\| \cdot \|d_j\|} = \frac{\sum_{k=1}^{n} w_{ik} w_{jk}}{\sqrt{\sum_{k=1}^{n} w_{ik}^2 \sum_{k=1}^{n} w_{jk}^2}}$$

Applying the VSM to document classification is very simple: each descriptor is associated to a vector in the space. Then, when classifying a new document *d*, descriptors are ranked in decreasing order of similarity *sim(d, desc$_i$)* for all descriptors *desc$_i$* in the thesaurus.

The vectors can be built for each descriptor in the following two ways:

- **VSM with independent classes:** we associate to each descriptor the terms of its own descriptor, and the ones belonging to its associated non-descriptors. Relationships between classes are not taken into account.
- **Hierarchical VSM:** each descriptor is represented by the vector containing its own terms, the terms of its associated non-descriptors, and the vectors of its broader term are added recursively until a top term is found[3]. Hierarchical VSM is supposedly more accurate as BT/NT relationships are used.

This approach is very similar to the one used in Adami et al. (2005) and is adapted to the case of thesauri.

Proposed Research Lines

This area has not been studied in depth and is therefore a promising field. We enumerate several research proposals directed at students in this subarea of text mining.

1. **Building several test collections.** As mentioned previously, official thesauri are used in many institutions. It would be interesting to collect a considerable number of manually indexed documents for a certain thesaurus into a test collection, together with the version of the thesaurus used to classify them.
2. **Development of new models.** This is of course the central point of research in this area: several classifiers are needed for experiments, improvements, etc. Three approaches can be considered here:
   a. *Clustering (i.e. unsupervised) approach*: a classifier which only uses thesaurus information to associate descriptors to documents.
   b. *"Pure" machine learning approach:* given a set of classified documents, building classifiers without using any information from the thesaurus. This task implies adapting classical text categorization methods to deal with some of the previously mentioned difficulties (high dimensionality and unbalanced data).
   c. "*Mixed" approach:* how to combine the "expert knowledge" arising from the thesaurus, and a good machine learning classifier to build a better automatic indexer.

3. **Evaluation of the results.** Finding the best way to evaluate results is also a very important question. As explained before, traditional precision-recall metrics (and derivatives) do not consider the hierarchical relationships between classes. A more precise measure should take this structure into account.
4. **Feature selection models for text.** We have the following questions. On the one hand, which parts of documents are relevant for automatic indexing? Do we need only the title of an article? Is an abstract also needed? The entire body? On the other hand, how can classical term selection methods be adapted to the case of thesaurus indexing?
5. **Feature selection for the thesaurus.** It is also important to find which data from the thesaurus helps our task. Are all the relationships equally useful? Should various relationships be eliminated? Should only certain thesaurus terms be selected? How should non-descriptors be considered?

REFERENCES

Adami, G., Avesani, P., & Sona, D. (2005). Clustering documents into a Web directory for bootstrapping a supervised classification. *Data Knowledge and Engineering*, 54(3), 301-325.

Berry, M. (Ed.). (2003). *Survey of text mining: clustering, classification, and retrieval*. New York: Springer.

de Campos, L. M., Fernández-Luna, J. M., Huete, J. F., & Romero, A. E. (2007). Automatic indexing from a thesaurus using Bayesian networks: application to the classification of parliamentary initiatives. *Proceedings of the 9th European Conference on Symbolic and Quantitative Approaches to Reasoning With Uncertainty, (ECSQARU 2007). Lecture Notes in Computer Science*, 4724, 865-877.

DESY Library (1996), *The high energy physics index keywords*. Retrieved December 1, 2007 from http://www-library.desy.de/schlagw2.html

Dumais, S. & Chen, H. (2000), Hierarchical classification of Web content. *Proceedings of the ACM SIGIR Conference (SIGIR 2000)*, 256-263.

Dumais, S. T., Platt J. C., Heckerman D., & Sahami M. (1998), Inductive Learning Algorithms and Representations for Text Categorization. *Proceedings of the 7th International Conference on Information and Knowledge Management (CIKM'98).* 148-155.

FAO World Agricultural Information Center (1998), *AGROVOC, multilingual agricultural thesaurus*. Retrieved December 1, 2007, from http://www.fao.org/scripts/agrovoc/frame.htm

Golub, K. (2006). Automated subject classification of textual Web documents, *Journal of Documentation*, 62, 350-371.

Hersh, W. R. (2003). *Information Retrieval: A Health and Biomedical Perspective, 2nd Edition*, New York: Springer.

Hersh, W. R., Buckley, C., Leone, T. J., & Hickman, D. (1994). Ohsumed: an interactive retrieval evaluation and new large test collection for research. *Proceedings of the ACM SIGIR Conference (SIGIR'94).*

Miles, A., & Brickley, D. (2005). *SKOS core guide, W3C working draft*. Retrieved December 1, 2007 from http://www.w3.org/TR/2005/WD-swbp-skos-core-guide-20051102

National Agriculture Library. (2002). *NAL Thesaurus*. Retrieved December 1, 2007 from http://agclass.nal.usda.gov/agt/agt.shtml.

National Library of Medicine. (1986). *Medical Subject Headings*. Bethesda, Maryland, USA.

Marjorie, H., & Hainebach, R. (1996). Multilingual Machine Indexing. Proceedings of the 9th International Conference on New Information Technology. Retrieved December 1, 2007, from http://joan.simmons.edu/~chen/nit/NIT'96/96-105-Hava.html

Maron, M. E. (1961). Automatic indexing: an experimental inquiry. *Journal of the ACM*, 8, 404-417.

Medelyan, O. (2005). Automatic keyphrase indexing with a domain-specific thesaurus. University of Fribourg, MSc. Thesis.

Medelyan, O., & Witten, I. H. (2006). Thesaurus based automatic keyphrase indexing. *Proceedings of the Joint Conference in Digital Libraries 2006 (JCDL 2006)*, 296-297.

Medelyan, O., & Witten, I. H. (2005). Thesaurus-based index term extraction for agricultural documents, Proceedings of the 6th Agricultural ontology service workshop at EFITA/WCCA 2005.

Montejo-Ráez, A. (2005), *Automatic text categorization of documents in the high energy physics domain*. PhD. Thesis, Universidad de Granada.

Montejo-Ráez, A. (2002), Towards conceptual indexing using automatic assignment of descriptors. Proceedings of the 2nd International Conference on Adaptive Hypermedia and Adaptive Web.

Moskovitch, R., Cohen-Kashi, S., Dror, U., Levy, I., Maimon, A., & Shahar, Y. (2006). *Multiple hierarchical classification of Free-Text Clinical Guidelines*, Artificial Intelligence in Medicine, 37, 177-190.

Office for Official Publications of the European Communities (1995), *EUROVOC. thesaurus Eurovoc - volume 2: subject-oriented version. Ed. 3/English language. Annex to the index of the Official Journal of the EC*. Retrieved December 1, 2007 from http://eur-lex.europa.eu/en/index.htm

van Rijsbergen, C. J. (1979). *Information retrieval (Second edition)*, London: Butter Worths.

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11), 613-620.

Sebastiani, F. (2002). Machine learning in automated text categorization, *ACM Computing Surveys*, 34.

Sebastiani, F., & Gabrilovich, E., *Bibliography on Automated Text Categorization*,. Retrieved December 1 from http://liinwww.ira.uka.de/bibliography/Ai/automated.text.categorization.html

Shannon, C. E. (1948). A mathematical theory of communication, *AT&T Technical Journal*, 27, 1948. Reprinted in: 2001, *Mobile Computing and Communications Review* 5(1), 3-55.

Steinberger, R. (2000). Using thesauri for information extraction and for the visualisation of multilingual document collections. Proceedings of the workshop on ontologies and lexical knowledge bases (OntoLex2000), 130-141.

Steinberger, R., Pouliquen B., & Hagman, J. (2002). Cross-lingual document similarity calculation using the multilingual thesaurus Eurovoc. *Proceedings of the Third Computational Linguistics and Intelligent Text Processing, CICLing'2002. Lecture Notes on Computer Science* 2276, 415-424.

Steinberger, R., Pouliquen B., & Ignat, C. (2003). Automatic annotation of multilingual text collections with a conceptual thesaurus, *Proceedings of the Workshop in Ontologies and Information Extraction (EUROLAN2003).*

Steinberger, R. (2001). Cross-lingual keyword assignment. *Proceedings. of the XVII Conference of the Spanish Society for Natural Language Processing (SEPLN2001).* 273-280.

Sun, A., Ee-Peng, L., & Wee-Keong, Ng. (2003), Performance measurement framework for hierarchical text classification. Journal of the American Society for Information Science and Technology, 54(11):1014-1028.

KEY TERMS

Descriptor: A word or a list of words that represents a concept without ambiguity. It can be used to retrieve documents in an information system, for instance a catalog or a search engine.

Document Indexing: Is the act of describing a document by index terms to indicate, with that metadata, what the document is about or to summarize its content. The index terms are often selected from some form of controlled vocabulary, e.g. a thesaurus.

Supervised Classification: Is a machine learning technique for creating a function from training data. The training data consist of pairs of input objects (typically vectors), and desired outputs (categories). The task of the supervised learner is to predict the value of the function for any valid input object after having seen a number of training examples (i.e. pairs of input and target output). To achieve this, the learner has to generalize from the presented data to unseen situations in a "reasonable" way.

Text Categorization: Is the task of assigning an electronic document to one or more categories, based on its contents. If only one category is assigned, we refer to that as single-label categorization. If several categories can be assigned to the document, we are dealing with multi-label categorization.

Thesaurus: A thesaurus is a list of every important descriptors in a given domain of knowledge; and, for each descriptor, the set of descriptor related with it.

Vector Space Model: Is an algebraic model for representing documents (not only text) as vectors of identifiers, such as, for example, index terms. It is used in information filtering, information retrieval, indexing and relevancy rankings. Its first use was in the SMART Information Retrieval System.

Unsupervised Classification: Is a machine learning technique where a model is fit to observations. In this case there is no a priori known output (as in supervised classification). In unsupervised classification, a data set of input objects is partitioned into different groups or clusters, so that the objects in each group share some common trait, e.g. proximity according to some defined distance measure.

ENDNOTES

[1] This work has been jointly supported by the Spanish Ministerio de Educación y Ciencia, and Junta de Andalucía, under projects TIN2005-02516 and TIC-276, respectively.

[2] It is also sometimes called a *concept* or an *index term*.

[3] We can consider this model to be one where every descriptor *contains* its own information, and all the associated information recursively contained by its broader term. A descriptor is therefore the *specialization* of its broader term.