

Categorisation Specification

Project ref. no.	LE4-8303
Project title	EUROSEARCH

Deliverable status	Restricted
Contractual date of delivery	Month 3
Actual date of delivery	Month 3
Deliverable number	D 4.1
Deliverable Title	Categorisation Specification
Type	Specification
Status & version	Final 1.0
Number of pages	17
WP contributing to the deliverable	WP 4
WP / Task responsible	UNIDO
Authors	Norbert Fuhr, Norbert Gövert, Mounia Lalmas, Fabrizio Sebastiani
EC Project Officer	Marina Manzoni
Keywords	categorisation, classification, probabilistic indexing, category description
Abstract	WP 4 deals with automatic categorisation of web documents. This deliverable describes two different categorisation tasks and gives a specification of methods used to perform these tasks. Categorisation is based on a description oriented approach to document indexing.

Summary

Work Package 4 (WP 4) is the automatic categorisation of web documents. Its objective is to implement a categorisation tool aimed at automatically generating web catalogues and enhancing document querying by restricting queries to specific categories. Deliverable D 4.1 is the *Categorisation Specification* which describes the methods that will be used to perform the tasks of WP 4.

The approach will be based on an automatic textual analysis of web documents. Weighted term features will be associated to documents. Their determination will be based on a probabilistic framework.

The weighted term features will be used to categorise web pages. The categories will be derived from a test-bed of documents from the *Computer and Internet Yahoo!* category.

Our categorisation tool will perform two tasks: the automatic classification of new documents into appropriate categories, and the querying of documents that belong to categories. To pursue the second task, a category description database will be generated.

Contents

1	Introduction	4
2	Architecture	5
3	Test-bed creation	6
3.1	Spidering of documents	7
3.2	Document normalisation	7
3.3	Statistics	8
3.4	Some issues	8
4	Document indexing	9
4.1	Term extraction step	9
4.2	Description step	10
4.3	Decision step	11
5	Category description generation	12
6	Applying classification methods	14
6.1	Document-centred categorisation	14
6.2	Category-centred categorisation	14
7	Conclusions	15

1 Introduction

Work Package 4 (WP 4) is the automatic categorisation of web documents. Its objective is to implement a categorisation tool aimed at automatically generating web catalogues and enhancing document querying by restricting queries to specific categories for the EuroSearch federation of search engines. The categories serve to partition the web space into subject-specific categories.

Although developing a local category catalogue for each search engine involved in the federation would better reflect each local culture, simply linking together existing local catalogues would not produce a catalogue capable to compete with the most popular US catalogues. Furthermore, a traditional catalogue is very expensive to build because it requires a large number of trained staff with a diversified expertise. One way to overcome these two problems is to build a document categorisation tool that, used on different contexts, allows to automatically build catalogues that reflect the local culture. The aim of WP 4 is to perform such a task.

The categorisation approach will be grounded on an automatic textual analysis of web documents associating weighted term features to documents. The determination of the weighted term features will be based on the probabilistic approach developed in [Fuhr & Buckley 91]. This strategy is a kind of long-term learning process that collects feedback from all information retrieval sessions, thus increasing the size of the learning sample over time. As a result, the estimation of the weights can be improved.

The weighted term features will be used to categorise web pages according to their content. The categories will be derived from a test-bed of documents from the *Computer and Internet* Yahoo! category. We use a test-bed containing English texts for evaluation purpose. It is important that we evaluate our categorisation tool, and that we compare it with other categorisation tools. The comparison is often only possible with tools dealing with English texts. However, since our approach is fully automatic, it is portable to the various languages involved in the federation.

The weighted term feature indexing will allow for two main tasks: the automatic classification of new documents into appropriate categories, and the determination of documents that belong to given categories. To pursue the second task, a category description database will be generated defining categories on the basis of weighted terms.

Deliverable D 4.1 is the *Categorisation Specification* which describes the methods that will be used to perform the tasks of WP 4. Its aim is to identify the indexing technology and classification tools. Its main activities are:

- determine the overall architecture of the categorisation task (section 2).
- define the domain and the document format for the test bed (section 3)
- specify the indexing approach (section 4)
- generate the category descriptions (section 5)
- specify the classification tools (section 6).

2 Architecture

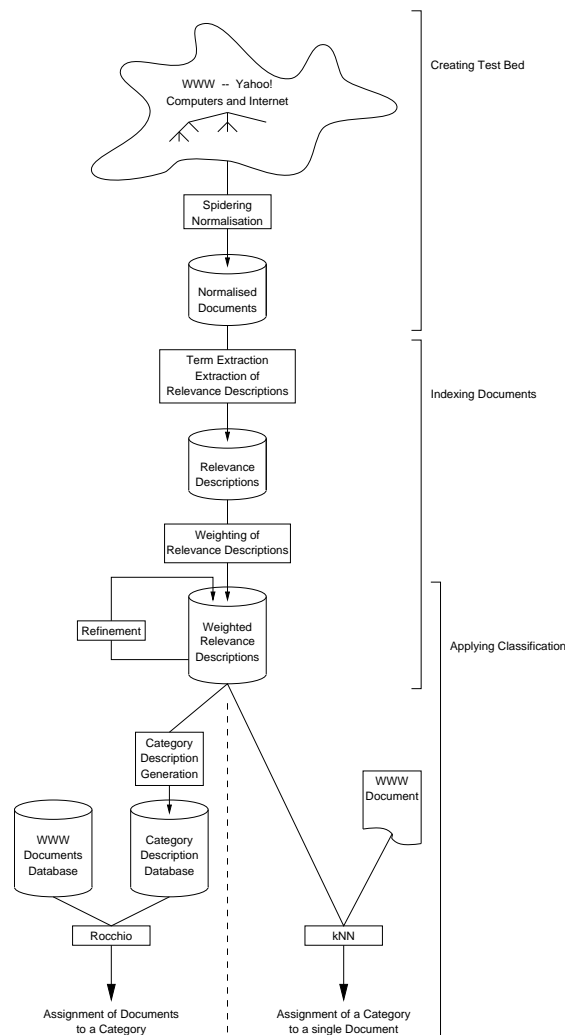


Figure 1: Architecture of the categorisation task

The overall architecture of the categorisation task (see figure 1) is split in three parts:

1. **Creating the test-bed** The categorisation of documents requires a test-bed of pre-categorised documents. For this purpose, the *Computer and Internet* category of the Yahoo! catalogue will be used. The first task will be the *spidering* of documents (section 3.1), and the second task will be the *normalisation* of documents necessary to handle the various structures of web documents (section 3.2).
2. **Indexing the document** Terms representing documents will be first extracted from the normalised document database (section 4.1). Document indexing will be defined in terms of term-document pairs, referred to as *relevance description*, which will be derived (section 4.2). The outcome will be a database of extracted features assigned to document-term pairs. Probabilistic weights will then be assigned to the terms (section 4.3) on the basis of the relevance descriptions. The weights will be refined through a learning process thus enhancing document indexing. Typical refinements consist of restricting the term space to discriminatory terms and adjusting weights. The outcome will be a database of documents indexed by weighted terms.
3. **Applying the classification** On the basis of the indexed document database, classification methods for the two categorisation tasks will be applied. The kNN method (section 6.1) will be used to automatically classify a new document. Inputs for kNN are the indexed document database, and the web documents which are to be classified. The Rocchio method (section 6.2) will be used to assign the best representative documents to some given categories. In this case, the indexed document database cannot be used as input, so each category will be described by a vector of weighted terms which will be derived from the indexed document database. The category description database along with a web document database will be used as inputs to the Rocchio method.

3 Test-bed creation

The creation of the test-bed requires two steps, the *spidering* of documents (section 3.1) and the document *normalisation* (section 3.2). Some statistics are given in section 3.3 and some issues are discussed in section 3.4.

3.1 Spidering of documents

Documents from the from Yahoo!'s *Computers and Internet* catalogue will be spidered. The process, in addition to the references of the documents themselves, will record the followings:

- relationship between documents and assigned categories;
- relationship between super and sub-categories; and
- (symbolic) linking between same categories in different parts of the sub-hierarchy.

In addition, the documents referenced by the documents directly referenced by Yahoo! categories will be spidered, and the corresponding anchors will be retained. This approach is necessary because the documents directly referenced by Yahoo! categories often are only entry points to a group of documents, that is a list of anchors with very little text (e.g., the document ACM/SIGLINK in the category *Computer and Internet: Multimedia: Hypermedia*). Indexing them on this basis will be ineffective (yielding low recall). The depth of the hierarchy (among the documents) will be of one so that to maintain high precision.

3.2 Document normalisation

Web documents vary immensely, so they must be normalised. The normalisation will yield the parts of the documents that will be considered to determine their content and hence categorising them. Our normalisation approach will be as follows:

1. We will consider only textual data. Therefore, images, speech, java scripts and applets, forms etc will be ignored for indexing purpose.
2. As discussed in the previous section, a Yahoo! referenced document may have limited textual data. A document referred by a Yahoo! category (*depth zero* documents) will be indexed on the basis of that document and the documents it links to (*depth one* documents). Such techniques proved successful in other work [Dunlop & Van Rijsbergen 93].
3. To restrict our test-bed, only referred documents that are on the same web site of the document referred by Yahoo! categories will be considered. This is to ensure high precision.

4. We will consider the following discourse: title, body, heading, first or last paragraph, etc. This can be obtained from the language HTML, upon which web documents are based. This will enable us, for example, to capture that in web documents, content bearing terms often appear at the top of the documents.

3.3 Statistics

The category *Computers and Internet* in Yahoo! contains 2533 unique sub categories, which are distributed over 7 levels:

level	categories
1	32
2	430
3	683
4	784
5	468
6	166
7	19

In addition there are 989 symbolic links: categories which appear at least twice in different places (these are counted once in the table above).

There are about 25500 documents referenced within the *Computers and Internet* catalogue (not counted are entries in those categories which are outside the *Computers and Internet* class).

3.4 Some issues

While investigating the creation of the test-bed from the Yahoo! site, some issues were raised:

- there is an average of 10 references per Yahoo! category. This may have an impact because the training set must be big enough to be representative. However, our approach will not be specific to the selected categories, and can easily be applied to any category division (for example, if sub-categories must be merged into super-categories).
- some levels of the Yahoo! classification contain both documents (references) and sub-categories. These documents will be part of the test-bed domain. As for above, our approach can be applied to any category level (whether higher or lower in the hierarchy).

- the Yahoo! catalogue cannot be represented by a tree; many of the categories are just symbolic links to other categories in the Yahoo! catalogue. The correct representation of the Yahoo! catalogue is a directed acyclic graph. Our approach will also handle this without problem.

The output of the spidering and normalisation processes will consist of a database of normalised documents with the following format:

- document identifier;
- textual data (with HTML markup);
- categories associated to the documents; and
- for depth zero documents (those directly referenced by a Yahoo! category), links to referenced documents (depth one documents).

4 Document indexing

To assign categories to documents (or vice versa) one has to find a suitable representation of the documents. This requires the indexing of documents. We will adopt a three-steps process: In the *term extraction* step (section 4.1), terms will be extracted. In the *description* step (section 4.2), relevance descriptions for term-document pairs will be constructed. In the *decision* step (section 4.3) probabilistic weights will be assigned to indexing terms on the basis of the relevance descriptions data.

4.1 Term extraction step

The term space will consist of single words and phrases. We will extract the terms from the document structure, i. e. by parsing the HTML source of the documents. Standard knowledge extraction methods for text will be used.

Removal of HTML markup will be first performed. Then documents will be filtered from stop words [Rijsbergen 79] and stemmed [Porter 80]. The outcome will be a list of single words to be used in the indexing vocabulary. The identification of phrases will be done by applying a part-of speech tagger and a phrase extractor. A robust lexicon based parser will be used for this task [Mikheev & Finch 97].

Other types of terms will be derived from the HTML markup. URLs can be used as well as the content of various tag attributes. URLs can provide

contextual information such as academic vs. commercial documents (e. g., “uni”, “ac”, “edu” in German, British, and American sites, respectively) or geography (e. g., “de” for Germany, “it” for Italy, etc.). The tag attribute ALT is a description of the picture where the SRC attribute in an IMG tag points to. Although images are not indexed, this information is useful for indexing the documents.

4.2 Description step

The description step consists of the construction of *relevance descriptions*. A relevance description for a term-document pair comprises a set of attributes which are considered to be important for the task of assigning weights to terms with respect to the documents. A relevance description $x(t, d)$ contains values of attributes of term t in document d . Attributes which will be considered are:

- dictionary information about t , e. g., inverse document frequency;
- type of term t , e. g., URL, within tag vs. within text term, phrase;
- parameters describing d , e. g., document length or number of terms in the document.
- information about the form of occurrence of t in d , e. g., discourse of d that contains t (title; main body; precise location such as beginning or end of document; factual data such as date, author), within-document frequency, or if t is a phrase the word distance between the first and the last component of t ;
- whether t is located in d or documents referenced by d (depth one documents);
- attributes of t in the documents referred by d ;
- other features such as context (academic vs. business), tag-based information, geography, etc.

The outcome of the description step process will consist of a database of triplets of document identifier, term and their associated relevance description. The relevance description format will be that of a n -dimension vector, where each dimension corresponds to an attribute (e. g., type of term, term frequency information, document parameters, etc.). The dimension entry will be the value associated to the attribute and n will be the number of considered attributes.

4.3 Decision step

In the decision step, documents will be indexed by *weighted terms*. The indexing process will be based on a *probabilistic information retrieval* approach using relevance data. The basic idea is to use long-term learning of indexing weights from previous information retrieval sessions in order to estimate the term weights.

We will use the probabilistic method proposed in [Fuhr & Buckley 91]. The goal is to determine term weights using a probability $P(R|x(t, d))$. This represents the probability that a document will be judged relevant to an arbitrary category given that one of the document's index terms which is also used to specify the category has the relevance description x . The advantages of this approach are numerous:

1. document-term pairs with different documents or terms can be mapped to the same relevance description (the same x). Therefore, the amount of relevance data that is available for the estimation of a specific indexing weight is not dependent on the number of categories, or documents for which we have relevance assessment. This overcomes a serious limitation encountered by many other probabilistic indexing approaches.
2. term weighting scheme can be derived for various forms of representation, for example:
 - (a) attributes specific to web documents can be explicitly taken account (e. g., document structure, tag information, etc).
 - (b) different types of terms can be manipulated (e. g., standard words, phrases, URLs, etc).
 - (c) the heterogeneous nature of web documents can be captured (text data, list of anchors, etc.)
3. the approach can be applied with little effort to web documents written in other languages than English. Obviously, we would need stemmers and stop word lists specific to other languages, but most of the other aspects of the indexing process (in particular in the decision step) are language-independent. This is crucial in this work since the categorisation tool will be used by the federation, where various European languages are involved.

The estimation of $P(R|x(t, d))$ will be performed using probabilistic regression methods which have shown to be successful (see for example [Fuhr 93]). We will generate the training sample from classified documents. The initial weights will be the standard $tf \cdot idf$ which will be used for generating the training sample.

The outcome of the decision step process will consist of a database of pairs of document identifiers and m -dimension term vector. Each dimension will correspond to a term and the value associated to the dimension will be the probabilistic term weight. m will be the number of terms considered in the indexing.

5 Category description generation

The approach to document categorisation we will follow is based on representing both categories and documents as vectors of weighted terms, so that the appropriateness of classifying document d under category c may be established by measuring the degree of similarity, or estimated probability of relevance, of d to c . We can then exploit well-developed techniques used in the standard information retrieval task, in which the appropriateness of retrieving document d as a result of query q may be established by measuring the degree of similarity, or estimated probability of relevance, of d to q .

Two categorisation tasks will be tackled in parallel, namely:

1. *document-centred categorisation*, in which, starting from document d , the most appropriate category (or categories) under which to categorise d are individuated;
2. *category-centred categorisation*, in which, starting from category c , the documents most relevant to c (e. g. those whose estimated probability of relevance exceeds a fixed threshold, or the n documents with the highest estimated probability of relevance) are individuated. In this task, a document may be categorised under one or several categories, depending whether it is considered relevant to one or multiple topics, or also under no categories at all.

Task 1 may be carried out without actually building a description for the category of interest, as the vector of weighted terms representing document d is compared with the vectors of weighted terms representing the documents in the test-bed. Given that the categories to which these latter belong are known, the category (or categories) under which document d should be categorised may be individuated by considering the categories under which the documents most similar to d had been categorised (see section 6.1).

In Task 2, instead, the term vectors representing documents in the test-bed belonging to category c cannot be used directly. A *description* of a category is needed, against which individual documents to be categorised are to be matched. Category descriptions are not available, so they must be

automatically built. In our approach, the description of each category c will automatically be extracted from an analysis of the test-bed documents that are categorised under c .

The extraction task will require a *term space restriction* (also known as *feature selection*) to be performed: this means that every category c will be represented as a vector of only those terms that are most significant, or specific, to c . The terms describing c will be selected from the *positive vocabulary*, i.e. the set of all terms appearing in the test-bed documents filed under c . This selection will be based on an automatic process that determines, for each such term t , the value of a suitable *feature selection metric*.

Basically, feature selection metrics tend to emphasise, each in its own way, those terms whose *within-category inverse document frequency* is substantially smaller than their *within-collection inverse document frequency*. This criterion is based on the intuition that terms specific to a given category occur more frequently in documents categorised under it than in “generic” documents. As a feature selection metric, we will use a derivation of the χ^2 method [Schütze et al. 95], for which the selection metric for term t is increased proportionally to the number of relevant documents containing t and irrelevant documents not containing t , and decreased proportionally to the number of irrelevant documents containing t and relevant documents not containing t .

Once the set of terms that will represent c have been chosen, their probabilistic weights have to be determined; the weight of term t will represent the relative importance of the associated term in describing the category. The weights must be such that each document d that was filed under c should actually be deemed to belong to c by the similarity function that we will use in categorisation, and each document d that was not filed under c should actually be deemed not to belong to c ; in other words, the category description should categorise *perfectly* the documents from the test-bed. For this purpose, a perceptron learning algorithm, that has already been successfully experimented [Ng et al. 97], will be used. The algorithm starts from a random assignment of weights to terms and incrementally refines them by using the weights contained in vectors of misclassified documents; the process iteratively converges, up to the point in which there are no more such documents.

The outcome of the category description generation process will thus consist of a database of pairs (category, term vector), where each element of the vector will consist of a pair (term, weight).

6 Applying classification methods

Using both the probabilistic indexing and the category description, documents can then be classified according to some given categories. There are two classification tasks. The first one is that given a document to identify its category (section 6.1). The second task is to search through a database the documents that satisfy best a given category (section 6.2). We will use two distinct approaches to carry out the two tasks, respectively, kNN [Yang 94] and Rocchio methods [Rocchio 71]. This decision is based on a survey/comparison paper [Yang 97] which allowed us to identify what was best for fulfilling each task.

6.1 Document-centred categorisation

Document-centred categorisation consists of assigning categories to a document. This task will be performed using kNN or k-nearest neighbour classifier which works as follows. Given an arbitrary document, the method ranks its nearest neighbour among training documents (the test-bed in our case), and uses the categories of the k top-ranked documents to predict the categories of a new document. The similarity score of each (k) neighbour document is used as a weight of its categories, and the sum of category weights over the k nearest neighbour are used for category ranking.

In [Yang 97] kNN was shown to be effective for the task of classifying single documents. The advantage of this method is that it is both simple and computationally efficient. kNN was also shown effective when the category space increases.

6.2 Category-centred categorisation

The category-centred categorisation consists of assigning documents to a category. To search documents from a given collection that represent a category best, the Rocchio algorithm will be used. The Rocchio method [Rocchio 71] is a vector space model for classification. For each given category the training set of documents (i.e., the test-bed documents) is used to derive a prototype vector. The category ranking given a document is based on a similarity comparison between the document and the category vector.

7 Conclusions

This deliverable aim is the specification of the categorisation task. We have proposed an architecture showing the main components for carrying out the task. The three main steps are the following: creation of the test-bed upon which the task will be experimented and validated; specification of the indexing methods that will be used to represent web document content, including the generation of a category description database; and finally the classification algorithms that will be used to categorise documents, and to query documents using the categories.

References

- Dunlop, M.; Van Rijsbergen, C.** (1993). Hypermedia and Free Text Retrieval. *Information Processing and Management* 29(3), pages 287–298.
- Fuhr, N.; Buckley, C.** (1991). A Probabilistic Learning Approach for Document Indexing. *ACM Transactions on Information Systems* 9(3), pages 223–248.
- Fuhr, N.** (1993). Representations, Models and Abstractions in Probabilistic Information Retrieval. In: *Information and Classification. Concepts, Methods and Applications*, pages 259–267. Springer, Heidelberg et al.
- Mikheev, A.; Finch, S.** (1997). A Workbench for Finding Structure in Texts. In: *Proceedings of the 5th Conference on Applied Natural Language Processing*.
- Ng, H.-T.; Gog, W.-B.; Low, K.-L.** (1997). Feature Selection, Perceptron Learning, and a Usability Case Study for Text. In: Belkin, N. J.; Narasimhalu, A. D.; Willet, P. (eds.): *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 67–73. ACM, New York.
- Porter, M.** (1980). An Algorithm for Suffix Stripping. *Program* 14, pages 130–137.
- van Rijsbergen, C. J.** (1979). *Information Retrieval*. Butterworths, London, 2. edition.
- Rocchio, J.** (1971). Relevance Feedback in Information Retrieval. In: Salton, G. (ed.): *The SMART Retrieval System - Experiments in Automatic Document Processing*. Prentice Hall, Englewood, Cliffs, New Jersey.
- Schütze, H.; Pedersen, J. O.; Hull, D. A.** (1995). A Comparison of Classifiers and Document Representations for the Routing Problem. In: Fox, E.; Ingwersen, P.; Fidel, R. (eds.): *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 229–237. ACM, New York. ISBN 0-89791-714-6.
- Yang, Y.** (1994). Expert Network: Effective and Efficient Learning from Human Decisions in Text Categorisation and Retrieval. In: Croft, B. W.; van Rijsbergen, C. J. (eds.): *Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 13–22. Springer-Verlag, London, et al.

Yang, Y. (1997). *An Evaluation of Statistical Approaches to Text Categorization*. Technical Report CMU-CS-97-127, Computer Science Department, Carnegie Mellon University.