Project ref. no.	*LE4-8303*
Project title	**EUROSEARCH**

Deliverable status	*Confidential*
Contractual date of delivery	*End of project*
Actual date of delivery	*End of project*
Deliverable number	*D6.2*
Deliverable title	*Test Results*
Type	*RE*
Status & version	*Final*
Number of pages	*18*
WP contributing to the deliverable	*WP6*
WP / Task responsible	*EIT*
Author(s)	*Martin Braschler (EIT), Norbert Gövert (UniDo), Claus-Peter Klas (UniDo)*
EC Project Officer	*Yves Paternoster*
Keywords	*Test results, evaluation, multilingual searching component, categorisation tool*
Abstract (for dissemination)	*Deliverable 6.2 presents the test results from both the EIT Multilingual searching component and the UniDo categorisation tool. Both systems have been evaluated using automatic methods. After briefly presenting the methodology, we then give the results obtained, followed by a discussion. The results of the evaluation of both components are very encouraging.*

1. Table of contents

2. Executive Summary

Deliverable 6.2 presents the test results from both the Eurospider multilingual searching component and the University of Dortmund categorisation tool. The multilingual searching component has been tested by doing German • Italian cross-language searches, while the categorisation tool was evaluated using two test-beds consisting of pre-categorised documents from the Yahoo! catalogue and the German DINO-Online catalogue from the World Wide Web. Both test approaches are fully automatic. After briefly presenting the methodologies, we describe the test-beds and the results that were obtained. We then go on to discuss the results before we close with conclusions.

The Eurospider multilingual searching component was tailored for the Eurosearch federated web scenario. The development was focused to make the component usable for the short, general vocabulary queries that are common in such a scenario. We show that the system compares favourably against the monolingual case for such queries. We also discuss how the same methods were integrated into an expanded system, specifically adapted to use for participation in the TREC text retrieval conferences. The performance of this expanded system was very encouraging in comparison to the competition.

Two approaches for automatic Web document categorisation have been developed, implemented and evaluated. We show that both approaches perform better in terms of categorisation accuracy compared to two baselines.

3. Full Description

3.1 Introduction

As part of the Eurosearch project, new components for multilingual searching and an automatic categorisation tool have been developed. This deliverable reports on test results for these components.

In Section 3.2 we describe the evaluation of the multilingual searching component developed at Eurospider Information Technology AG.

Related deliverables include:

- D3.2 "Multilingual Components Specification", giving details on the inner mechanisms of the methods used in the translation component
- D3.3 "Integrated Translation Prototype", describing the prototype implementation of the translation component

Section 3.3 includes the evaluation of the automatic categorisation tool developed at University of Dortmund. Related deliverables include:

- D4.1 "Categorisation specification", giving a specification of the categorisation task and the architecture of the categorisation tool
- D4.2 "Probabilistic indexing and categorisation tool, intermediate prototype", describing methods used for automatic categorisation and the implementation of the first prototype of the tool
- D4.3 "Categorisation tool, final prototype", describing the implementation of the final prototype of the categorisation tool

A specification of evaluation methods used for both the multilingual component and the categorisation tool have been given in

- D6.1 "Test Specification"

3.2 The Multilingual Searching Component

3.2.1 Approach used

In Eurosearch, query translation is used to bridge the language gap when a user from one of the partner sites enters a query in a language that cannot directly be used for searching on the target collection. This query translation step takes place in the translation server, more specifically in the translation component. By using query translation, a "cross-language search" (using a query in a language different to the document language) is divided into two steps: first query translation, then a monolingual search.

Consequently, testing a multilingual searching component following this approach is very similar to testing a monolingual searching component, with the translation step added. We can therefore build on the well-known measures and methods developed for monolingual information retrieval evaluation[1].

The Eurospider multilingual search component uses similarity thesaurus technology to translate queries. For the time being, it offers translation to and from Italian. However, the component has been designed to be general enough to allow easy incorporation of further languages. Some tests with French and English have been made to prove the viability of such extensions. Since these language pairs are not part of the Eurosearch federation, they are however not the subject of this evaluation.

3.2.2 Similarity Thesauri

Similarity thesauri are constructed automatically from suitable training collections. This means the Eurospider multilingual search component uses no costly hand-tuned linguistic resources. A discussion of the implications of using automatically built resources can be found in deliverable 3.2. Generally speaking, the similarity thesaurus shows special strengths when used for domain-specific searches. However, it is more difficult to use it for more general document collections. Therefore, a lot of work during the Eurosearch project went into generalising the thesauri, in order to make them usable for a wide range of document collections.

Similarity thesauri contain for every thesaurus term a list of terms that are statistically similar. These similarities can be calculated from suitable training data. While similarity thesauri were originally developed for monolingual applications [Qiu:1995], the same techniques can be applied to multilingual settings as well, so long as multilingual training data is available. The thesaurus then contains for every head term in the source language a list of terms in the target language that are similar. The use of statistical methods to calculate these similarities means that the resulting thesaurus reflects the domain of the training data. We therefore applied various filtering runs on

[1] In this deliverable, we use the terms "search system/component and information retrieval system interchangeably.

the thesauri used for Eurosearch, to ensure that they cover a wide range of topics and are usable in the federated scenario of the project.

3.2.3 Information Retrieval evaluation

Since the multilingual searching component for Eurosearch essentially consists of query translation, followed by conventional monolingual searching, the methods and measures developed for evaluating classical Information Retrieval (IR) systems can be applied. The two most popular evaluation measures for automatic evaluation of IR systems are "*Precision*" and "*Recall*".

In order to calculate these measures, the result list returned by the system for a test query is examined. Precision and Recall are then defined as:

$$precision = \frac{number\_of\_relevant\_documents\_retrieved}{total\_number\_of\_documents\_retrieved}$$

$$recall = \frac{number\_of\_relevant\_documents\_retreved}{total\_number\_of\_relevant\_documents\_in\_collection}$$

where a document is considered "relevant" if it contains a good answer to the test query.

Obviously, it is desirable to maximise both precision and recall. These can be conflicting goals, however, since by adding additional search terms to enhance recall, the amount of irrelevant information typically also increases, thus diminishing precision. Depending on the application, users are more precision-oriented or more recall-oriented. A precision-oriented user is looking for an answer to a specific question, such as the date of the next elections. Usually, one relevant document is then enough to fulfil the information need. The user is not likely to be interested in more documents providing the same information. A recall-oriented user, however, is looking for as many relevant documents as possible. An example would be a user doing patent searches.

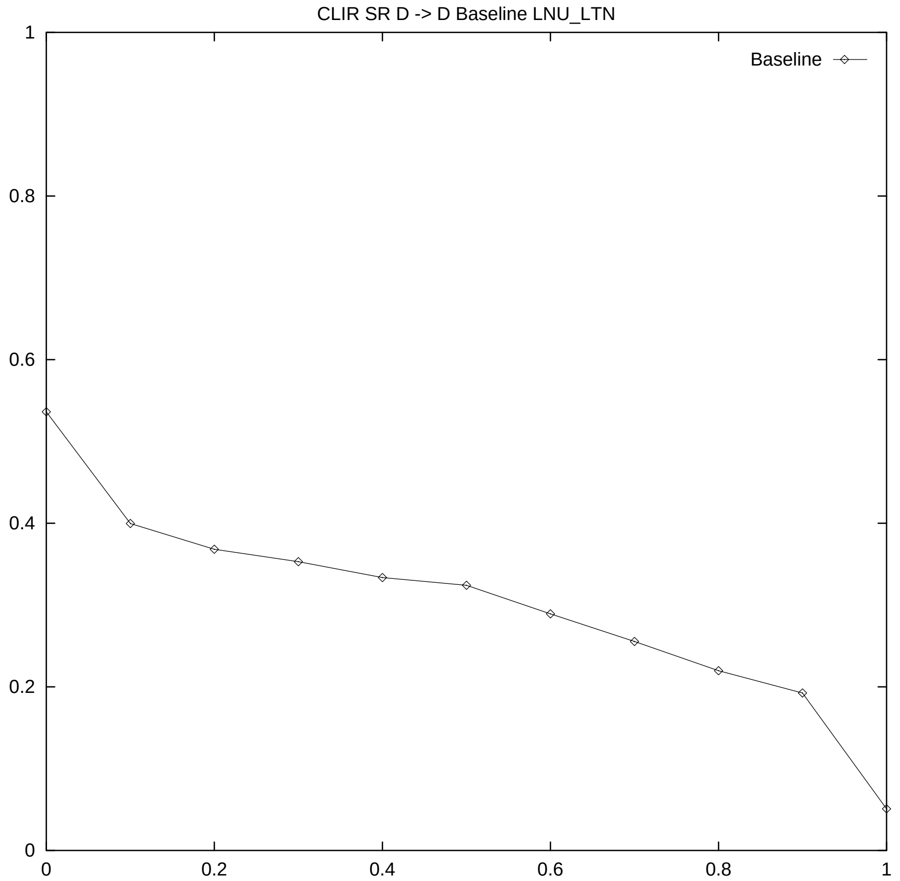
Most modern information retrieval systems[2] return their results as a ranked list. This allows to strike a balance between recall and precision: because the information that is most likely relevant to a search request is presented at the top of the list, a precision oriented user can inspect the top

[2] including the most popular web search engines

documents only. If good recall is essential, the user can continue down the list to find more relevant information.

Consequently, in most literature, precision/recall figures are presented as a continuous graph, showing the change of precision and recall depending on how many documents are inspected. The result is a curve, going from high precision and low recall to low precision and high recall.

Often, the result is also represented by a single figure, the so-called "average precision": precision is averaged over different levels of recall. The higher the average precision, the better the system performs.



A recall-precision graph. The y-axis gives the precision, while the x-axis shows different levels of recall.

3.2.4 The TREC test collection

To use the precision and recall measure, a test collection is needed, consisting of test documents *and* a set of test queries for which all relevant documents have been determined by a human assessor. This assessment by relevance is an extremely costly task for large collections. Therefore, in nearly all cases IR evaluation these days is done with pre-prepared test collections. This has the advantage that it is possible to test on a much larger document set than would be practical were the relevance assessment to be made specifically for a few test runs. However, it also means that the test data is not identical to the data ultimately being used in the system.

The most popular test collections for the past years have been the ones used in the TREC series of conferences (TREC=Text REtrieval Conference) [Voorhees/Harman:1999, Voorhees/-Harman:2000]. The TREC collections are large in comparison to most other test collections, which is important in evaluating today's large search systems. Since TREC is organized by the US National Institute of Standards and Technology (NIST), the data being used was originally only in English. However, with the growing importance of access to globally available data, TREC has evolved over the past years to include Spanish (now discontinued), Chinese, German, French and Italian texts.

Beginning in 1997 with TREC-6, a cross-language "sub-track" to TREC was introduced [Braschler/etal:1999A], a process in which Eurospider was directly involved. In connection with this extension to the TREC conferences, multilingual document collections were introduced. It is these collections we can use for testing of our translation component. Starting in 2000, these activities will become part of the DELOS network of excellence, funded by the Information Societies Technology programme of the European Commission Fifth Framework.

Table 1: Document collections used in the TREC Cross-Language track

English	AP news	1988-90	242,918 docs	750 MB
German	SDA news	1988-90	185,099 docs	330 MB
German	NZZ articles	1994	66,741 docs	200 MB
French	SDA news	1988-90	141,656 docs	250 MB
Italian	SDA news	1989-90	62,359 docs	90 MB

3.2.5 The test setup

Since we are testing a German • Italian translation component, we are using the TREC German and Italian test collections and queries.

Our test setup works by first running monolingual tests, i.e. German queries against German documents, and Italian queries against Italian documents. This gives us the baseline to compare our results to. Of course this is a very idealistic baseline, since most translations are ambiguous, therefore introducing additional uncertainty. However, it is a realistic baseline since it represent the quality of search results users have come to expect.

TREC uses so-called “topics” that describe the information need of a (virtual) user. It is then the task of the individual TREC participants to produce queries out of these topics, i.e. something that their search system can handle. It is important to note that this step has been done *automatically* in our test setup, therefore *not* tailoring the queries to unrealistically exploit special characteristics of the system. TREC topics contain various fields, making it possible to generate queries with different lengths. Usually, three different lengths are used with TREC experiments, so we adapted this convention for our tests. In the following, the three different generated query sets will be called short, medium-length and long queries[3].

Short queries usually consist of one to three query terms, medium-length of around 10 terms, and long queries of as many as 30-40 terms. This means that for Eurosearch, a project that is most concerned with a federated scenario of web search services, the short queries seem to have the most practical significance, since it is well known that most users of web search engines tend to enter very short queries. On one of the search services that Eurospider Information Technology hosts on the Internet[4], the average query length is between around 2.4 and 3.1 terms per query, depending on the query language (German, French and Italian). In some informal reports, similar results for some of the big web engines were suggested. In 1997, Excite reported during a panel session at the SIGIR conference that the average query length in their system was 2.2 terms (up from 1.5 terms in 1996).

We ran tests with both the TREC-7 and TREC-8 query set. TREC-7 was the first year that added Italian documents and queries. The TREC-8 relevance assessments are still very new (from mid-October 1999), so the analysis of the results is necessarily limited.

[3] short equals to the title field in TREC, medium-length equals to the title plus description fields in TREC, and long equals to all fields in TREC

[4] the search server for the decisions of the Swiss supreme court

3.2.6 Results with TREC-7 and TREC-8 queries:

We obtained the following results for the queries from TREC-7:

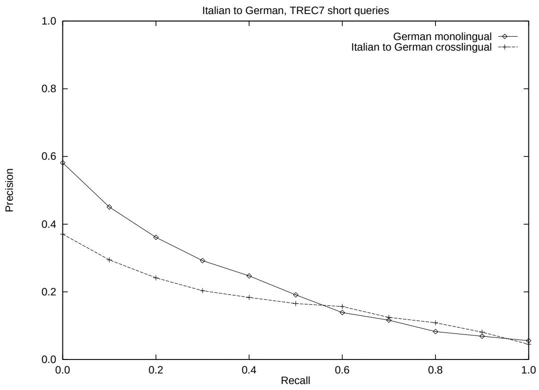
TREC-7	German monolingual	Italian to German	Percentage
long queries	0.3250	0.1762	54.2%
middle-length queries	0.2738	0.1536	56.1%
short queries	0.2172	0.1655	76.2%

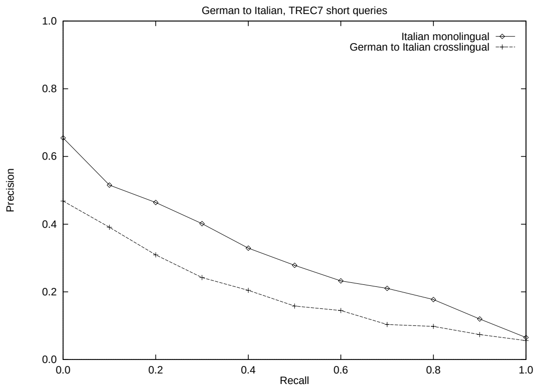
TREC-7	Italian monolingual	German to Italian	Percentage
long queries	0.4078	0.1629	39.9%
middle-length queries	0.3233	0.1670	51.2%
short queries	0.2928	0.1931	65.9%

For the queries from TREC-8, we obtained the following results:

TREC-8	German monolingual	Italian to German	Percentage
long queries	0.3072	0.1278	41.6%
middle-length queries	0.2543	0.1261	49.6%
short queries	0.1895	0.1038	54.8%

TREC-8	German monolingual	Italian to German	Percentage
long queries	0.4387	0.1271	29.0%
middle-length queries	0.3618	0.1430	39.5%
short queries	0.3219	0.1299	40.4%





The monolingual results are better the longer the query. This is what should be expected, since additional search terms help retrieve more good documents. However, this observation does not hold with the crosslingual case: longer queries also add more ambiguity. Therefore it is harder for the system to still produce a translated query which is focused on the original information need. We observe in our tests, that the three query lengths are performing essentially at equal level for the crosslingual case.

This means of course, that the crosslingual results compare most favourably against the monolingual results for the short queries. Note however, that this is clearly the most interesting case for the Eurosearch project. Our goal was to produce a translation server for a federated web environment. In this regard, the results from the TREC-7 short queries are very encouraging, especially the 76% performance we obtained for Italian to German.

The German to Italian results are somewhat lower. We think this is due to the higher absolute performance of these runs. We are not sure what makes the TREC-8 queries perform substantially worse. General trends we observed hold (longer queries better for monolingual, no specific preference for query length for the crosslingual case, short queries compare best), but in terms of relative performance we do not reach the performance of the TREC-7 set.

It is important to remember that since we wanted to test the web scenario for which the translation component in Eurosearch was developed, we did not enhance the system with features taking specific advantage of the characteristics of the test collection. In addition to the test runs for this deliverable, Eurospider incorporated the methods described in this report into a system geared for the test scenario used in TREC. The results of the official Eurospider participation to the TREC are not directly comparable to the results given in here, since in TREC, documents from multiple languages have to be searched simultaneously.

It is encouraging to note, however, that the Eurospider system performed very well at TREC-7, producing one of the best multilingual runs and comparing very favourably against machine translation based approaches. The system was also competitive at TREC-8, enforcing our belief that the technology used in the translation component is working well. This fact also makes us confident that the somwhat lower figures for the TREC-8 tests do not signify a general problem in the translation component. Full details on these TREC participations are in [Braschler/etal:1999B, Braschler/etal:2000].

3.3 Automatic Evaluation of Categorisation

3.3.1 Automatic Categorisation of Web Documents

The automatic categorisation of web documents is crucial for organising the huge amount of information available in the Internet. Within the Eurosearch project an automatic categorisation tool has been built.

One approach implemented for categorising Web documents is based on a probabilistic description-oriented indexing approach for Web documents. As opposed to traditional approaches for representing Web documents for the purpose of categorisation, the description-oriented approach captures the rich structure of web documents. The description-oriented approach is described in detail in Deliverable 4.3. Also in Deliverable 4.3 we presented experimental results, which show that this approach yields effective categorisation.

In addition the megadocument approach for categorisation has been devised and implemented. This new approach is described by Klas and Fuhr in [Klas/Fuhr:2000] and [Klas:1999]. In this approach, in a given catalogue schema, for each category the training documents are concatenated to form the so-called megadocument for the given category. The resulting megadocuments are indexed using standard information retrieval methods [Salton/Buckley:1988]. In order to classify a new document, the most similar megadocument with respect to the given document determines the category to be assigned.

In order to evaluate the effectiveness of our approaches, we compared it to a baseline made up of traditional methods for categorising documents. This baseline has been implemented within the intermediary prototype and is described in Deliverable 4.2.

3.3.2 Automated Evaluation of Categorisation Methods

For doing an automatic evaluation of categorisation methods, we chose a measure based on the standard measures for evaluating information retrieval methods, i. e. recall and precision (see

Section 3.2.3). We measure the precision in the first rank. Basically this is the percentage of documents being categorised correctly by a given classifier.

In the case that pre-categorised documents only show up once in a given categorisation schema, the measure `precision in the first rank' is equivalent to the measure `break-even-point between recall and precision', which is widely used in the context of evaluating categorisation methods (see e. g. Yang [Yang:1999]). In both test collection (described in the following section) we have almost the case that documents are categorised uniquely.

3.3.3 Test Collections

Automatic Evaluation of the effectiveness of categorisation methods requires a test collection of pre-categorised documents. Having such a test collection, traditional evaluation methods in information retrieval (i.e. recall and precision) can be applied, as described in Deliverables 4.3 and 6.1.

We chose two test collections of pre-categorised documents. One test collection consists of documents spidered from Yahoo!'s "Computers and Internet" catalogue (http://www.yahoo.com/). The creation of the Yahoo! test-bed has been described in Deliverable 4.1.

The second test collection consists of documents spidered from the German DINO-Online catalogue (http://www.dino.de/). The objective of choosing a German language catalogue is to demonstrate that the automatic categorisation tool can be extended easily in order to be applicable to other languages than English.

The following table shows statistical details of the two test-beds:

Feature	**Yahoo!**	**Dino**
Number of categories	2806	1211
Number of top-level categories	25	23
Number of documents	17710	55672
Average size of documents in bytes	127483	4006
Average number of terms per document	3684	107
Average number of distinct terms per document	721	69
Average maximum term frequency per document	149	6
Average number of non-root nodes per document	10	0
Size of term space	546527	284440

Both test-beds have been split into two disjoint document sets for training and testing. In both cases we used about 70% of the documents for training our categorisation tool while 30% of the documents were used for evaluating the tool.

3.3.4 Extending the Categorisation Tool for Multilinguality

The categorisation tool is extensible in order to cope with languages ther than English. For each language to be processed, two modules need to be added to the categorisation tool:

- elimination of stop words
- stemming, groundform reduction

Currently these modules are available for English and German.

3.3.5 Test Setup

Both test-beds have been split into two disjoint document sets for training and testing. In both cases we used about 70% of the documents for training our categorisation tool while 30% of the documents were used for evaluating the tool.

In a training phase our categorisation tool explores knowledge from the training documents in order to learn categorisation. Within the testing phase the test documents are fed into the categorisation tool; for each document a ranking of categories is returned, denoting towhich categories the system would assign the given test document. The decision rule taken for the evaluation described in this deliverable was to assign the document to the top-ranked category delivered by the system. In order to compute the precision in the first rank, we took this decision and compared it to the category which has been originally assigned in the Yahoo! or DINO-Online catalogue.

The following parameters have been used for our evaluation. The evaluation was done with respect to the top-level categories in the two test-beds. From each test-document the fifty most significant terms have been used in the categorisation process.

Experiments have been performed with respect to both approaches for automatic categorisation implemented in our categorisation tool. Two baselines were chosen to compare the effectiveness of our categorisation tool to. One baseline has been described in Deliverable 4.2 and is based on a document representation, which is standard in information retrieval and uses a tf x idf weighting of terms [Salton/Buckley:1988]. The second baseline is taken from a work from Chakrabarti et al. [Chakrabarti/etal:1998]. That second baseline is comparable to our setup in that the test-bed also has been taken from the Yahoo! catalogue and the evaluation has been done with respect to the top-level categories.

3.3.6 Results

The following table shows the results of the evaluation of our categorisation tool and the results obtained from the two baselines. Each row describes the result of one experiment. The first column gives a short description of the settings while the second column gives the precision in the first rank as a percentage of correct category assignments:

Experiment	**Precision in 1st rank**
Megadocument, Yahoo!	48.30%
Megadocument, DINO-Online	47.46%
Description-oriented approach, Yahoo!	36.50%
Baseline 1, Yahoo!	33.57%
[Chakrabarti/etal:1998], Yahoo!	32.00%

The results show that our two approaches to automatic categorisation perform noticeable better than both baseline implementations. The megadocument approach reaches a categorisation accuracy near 50% on both test-beds.

The probabilistic description-oriented approach does not perform as well as the megadocument approach. An analysis of the reasons is given in Deliverable 4.3 and in [Gövert/etal:1999].

4. Conclusions

The goals of the translation component in Eurosearch were two-fold:

1. Building a translation component that allows federated use to build a federation of European search engines, and
2. Building a translation component that is suitable for web-style searches, i.e. short, unstructured queries covering a wide range of topics.

Goal 1 is outside of the scope of this document. For details on how integration was achieved, please refer to the relevant deliverables (D3.3, D5.1).

This document is concerned with goal 2. During the Eurosearch project, the Eurospider similarity thesauri have been tuned to make them suitable for the general vocabulary of the web. The translation component built using these thesauri has been tested using the methodology and data of the TREC series of conferences. We observed that the component works best for short searches, which is the main application it will receive inside the Eurosearch federation. The component was shown to be easily extendible with new languages. We are encouraged by the test results we received, and an expanded system compared favourably at the TREC conferences.

Concerning the automatic categorisation tool we were able to show that our categorisation approaches both lead to enhanced precision with respect to categorisation, compared to two baselines. Also the tool was shown to be easily extendible with new languages. Currently it is implemented to cope with English and German.

5. References

Braschler/etal: 1999A

Braschler M., Krause, J., Peters, C., Schäuble, P. (1999). *Cross-Language Information Retrieval (CLIR) Track Overview.* In: Proceedings of the Seventh Text Retrieval Conference (TREC-7).

Braschler/etal.: 1999B

Braschler, M., Mateev, B., Mittendorf, E., Schäuble, P., Wechsler, M. (1999). *SPIDER Retrieval System at TREC-7.* In: Proceedings of the Seventh Text Retrieval Conference (TREC-7).

Braschler/etal.: 2000

Braschler, M., Kan, M.-Y., Klavans, J., Schäuble, P. (2000). *The Spider Retrieval System and the TREC-8 Cross-Language Track.* To appear.

Chakrabarti/etal:1998

Chakrabarti, S.; Dom, B.; Indyk, P. (1998). *Enhanced Hypertext Categorisation Using Hyperlinks.* In: Proceedings of SIGMOD 1998

Eurosearch consortium:

Deliverables D3.2, D3.3, D4.1, D4.2, D4.3, D6.1.

Gövert/etal:1999

Gövert, N.; Lalmas, M.; Fuhr, N. (1999). *A probabilistic description-oriented approach for categorising Web documents.* In: Proceedings of the 9th international conference on Information and knowledge management. **http://ls6-www.cs.uni-dortmund.de/ir/publications/1999/Goevert_etal:99.html**

Klas:1999

Klas, C.-P. (1999). *Ein neuer, effektiver Ansatz zur Kategorisierung von Web Dokumenten.* In: Heuer, A. (ed.). Proceedings ADI'99 (Agenten – Datenbanken – Information Retrieval).

http://ls6-www.cs.uni-dortmund.de/ir/publications/1999/Klas:99.html

Klas/Fuhr:2000

Klas, C.-P.; Fuhr, N. (2000). *A new Effective Approach for Categorising Web Documents.* Submitted.**http://ls6-www.cs.uni-dortmund.de/ir/publications/2000/Klas_Fuhr:00.html**

Qiu:1995

Qiu, Y. (1995). *Automatic Query Expansion Based On A Similarity Thesaurus.* In: PhD Thesis, Swiss Federal Institute of Technology (ETH).

Salton/Buckley:1988

Salton, G.; Buckley, C. (1988). *Term Weighting Approaches in Automatic Text Retrieval.* In: Information Processing and Management 24(5), pages 513-523.

Voorhees/Harman:1999

Voorhees, E.M., Harman, D.K. (1999). *Overview of the Seventh Text Retrieval Conference (TREC-7).* In: Proceedings of the Seventh Text Retrieval Conference (TREC-7).

Voorhees/Harman:2000

Voorhees, E.M., Harman, D.K. (2000). *Overview of the Eighth Text Retrieval Conference (TREC-8).* To appear.

Yang:1999

Yang, Y. (1999). *An Evaluation of Statistical Approaches to Text Categorisation.* In: Information Retrieval 1(1), pages 69-90.