

Weaving Intranet Relations – Managing Web Content

Ulrich Bohnacker, Lars Dehning, Jürgen Franke, Ingrid Renz, René Schneider

DaimlerChrysler AG
Research and Technology
P.O.Box 2360
89013 Ulm, Germany

{ulrich.bohnacker, lars.dehning, juergen.franke, ingrid.renz, rene.schneider}@daimlerchrysler.com

Abstract

We will give an overview of the WIR (Weaving Intranet Relations) system, a tool for offline computation and online retrieval of similar intranet documents. Text similarities are computed from a large collection of HTML-Documents and represented in a similarity matrix. With a click on a "What's Related?"-button, the user starts an intranet query for comparable documents with the full text as query input and receives a list of similar texts ranked by their corresponding similarity. The system is fully implemented and integrated into the intranet of a major company.

1 Motivation

Currently intranets play an important role in large companies and their impact as knowledge management tools on efficient business communication and administration will grow considerably in the future. Thus at an already early stage, the intranet collection of texts, images, videos etc. might turn into a confusing pell-mell. Very often people in large companies do not know that colleagues in another department have the know-how they need to solve their problems. In this case, queries for related documents in the intranet may be a big help for the processes of finding and the integration of documents. For effective retrieval and maintenance in these steadily growing intranets, considerable approaches have been developed (Agosti, Crestani, Melucci 1994, Allan 1995, Salton, Singhal, Buckley, Mitra 1996, Shin, Nam 1997), but further supporting tools are still needed. We developed a new tool: **Weaving Intranet Relations - WIR** which basically gives an innovative retrieval function. Additionally, it is able to support the organization and maintenance of the web content by suggesting new structures and links.

1.1 Retrieval

For any common user of huge text collections as the I*nets (Internet, intranet, extranet) the main work consists in finding the relevant information. This search usually consists in looking for some user-defined keywords. Every text which contains these words is presented to the user as a good candidate according to the inquiry. But the main problem is that the user must know which keywords the author wrote. Especially in new and innovative areas, this is a difficult task.

In short term queries, several natural language phenomena do not lead to satisfying results, e.g. whenever synonymy (a *monitor* in one text may be called a *display* in another document and *CRT* in another), homonymy (*Java* might be mentioned in a travel report of the executive chair or the name of a programming language) and polysemy (*table* might be a piece of furniture or an instance of assembling to eat) might play a role. These are major reasons why a full-text oriented search leads more often to better results and is especially preferable whenever the amount of the data collection is estimable and similarity computation is feasible within an acceptable time frame. Due to this reason we have chosen a deductive way of computing text similarity from a large collection of intranet documents.

Based on our WIR technology we developed a system that presents to a given text (i.e. query) thematically related documents not according to some words or concepts, but according to the whole given text. Here, the WIR system is started by a button named „What’s Related?“. Clicking this button is the only action the reader has to perform to get a list of thematically related texts.

1.2 Maintenance

Not only the reader/searcher/surfer but also the writer or manager needs better support for the intranet. Usually, the documents should be organized according to a given classification scheme and annotated with pre-defined index terms. This can be difficult since the same categorical or terminological structure may be differently understood and used. WIR is able to suggest appropriate categories and terms for any new text. Again, WIR computes thematically related sites – and their annotations are considered as possible specifications for the new document.

The same WIR technology serves to find duplicates or out-dated versions of the same content. This information about the web content can be used to remove sites, a task which will be essential in future.

At current state (February 2000), the intranet collection we are working with consists of about 40,000 documents after having increased from 9800 within two years. This remarkable growth of the collection underlines the significance that intranets do nowadays have in companies.

Fig. 1 illustrates the functionality of our WIR system:

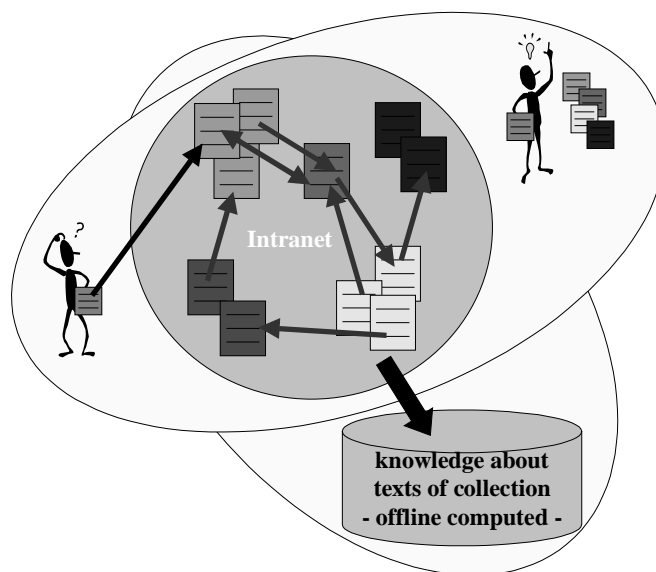


Figure 1: WIR – Weaving Intranet Relations: a tool for searching in and maintaining of intranets

2 System Overview: Offline Computation and Online Presentation

The core technology of WIR is the computation of thematically-based text similarities. These texts may belong together in various ways (multi-dimensionality). Their relation is mathematically calculated as similarity between vector representations of each text.

WIR integrates algorithms of corpus-based linguistics, information retrieval and statistical pattern recognition in order to compute the similarity of texts in a given electronic text collection like an intranet. The main characteristics of our approach is that no external knowledge (lexicon, terminology, concepts, etc.), but only collection-internal knowledge (i.e. statistics) is used. The system automatically adapts to any collection (unsupervised learning). Therefore the technology can be used for texts in any language and of any subject.

The system architecture of WIR is divided into an offline computation which calculates the relations of the texts and an online inspection of these relations given an inquiry (i.e. a text).

These two main parts the system is divided into are illustrated in fig. 2:

- offline computation of document similarities with the following stages: text normalization, feature selection, vector generation and similarity calculation. In a separate step, different similarity matrices that were calculated for different specified features and/or with different similarity measures are merged into a unique distance (similarity) matrix.
- online retrieval of similar documents through the call of a program that leads to the presentation of the ranked list of the corresponding documents according to their similarity values.

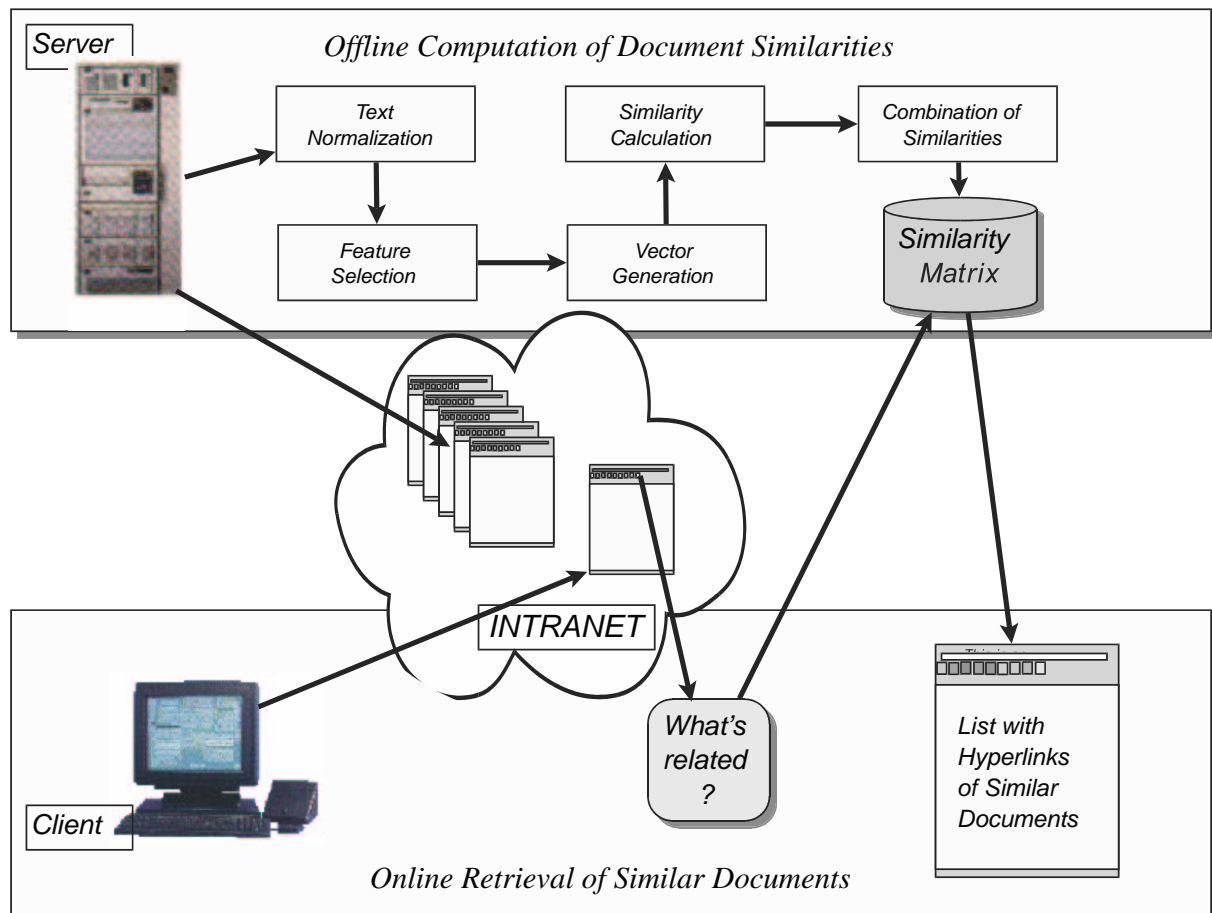


Figure 2: Offline computation of similarities and online presentation of search results

The first procedure, the calculation of similarities, works offline, because it is time expensive. All documents of the intranet are analyzed to get the similarities of all pairs of documents. Calculating the similarities works as follows: A text normalization module transforms the different formats of the real-world documents (in intranets we usually find HTML-files or documents of text processing applications, otherwise the documents consist of plain text files which may contain special characters). Feature selection collects all text elements for all documents in the given collection (here: the intranet). A text element is any word, but also any character string of given length (mostly 3 or 4, tri-gram or quad-gram). If we would integrate a morphological analyzer, any stem or morpheme could be used as text element. Out of these text elements, (statistically) qualified ones are picked as features and stored in a feature lexicon. This step drastically reduces the number of elements: from more than 150,000 down to less than 10,000. Vector generation builds a feature vector for each document by weighting the features of the feature lexicon. In the step of similarity calculation, the distance between these vectors is interpreted as similarity measure between the documents, a value that is stored into a similarity matrix. Since different algorithms and parameters can be used in any processing step, several similarity matrices can be computed which in a separate step are combined.

2.1 Offline: From Texts to Vectors

The task of weaving intranet relations can be seen as finding the most similar documents in an intranet or text collection. But the question is: On which characteristics should this relationship of textual similarity be computed? In order to characterize texts, typically documents are represented as vectors of words, see for example the popular vector representation for information retrieval (see Salton, McGill 1983).

We define word forms as the typical basic entity. Work in related fields like information retrieval, text categorization, information filtering or routing (presented at conferences like ACL-SIGIR - Special Interest Group Information Retrieval - or TREC - Text Retrieval Conference) has proven that this baseline can easily be improved by further procedures of feature generation and selection.

In languages like English without an elaborate system of morphological alternations, the selection of word forms often is a good basis for such tasks in information retrieval or text categorization. But in highly inflectional languages like German, generated features, i.e. shorter forms like word stems seem to be more useful.

Generally we distinguish between the shape of the features (full word forms, morphologically correct word stems, character strings of given length - n-grams, morphemes, phrases consisting of more than one item), the selection of features (all, only the most frequent ones or selected by different measures), the weighting of features(e.g. tf/idf), and the origin of features (from the whole text, from selected text passages, from other texts which are 'neighbors' in the given hyperlink structure or from external knowledge bases like index lists or thesauri).

For the first prototype, we explored different simple approaches, mainly from work in information retrieval but also from work in computational linguistics. Regarding the shape of features we computed full forms as baseline as well as character strings of fixed length (n-grams with $n = 3$ and 4). Both are computed on the basis of all texts in the given intranet or text collection. Thus, their number is extremely high (150,000 for full forms and more than 350,000 for quad-grams), too high for the following step of computing the text similarities. Therefore, selection procedures which choose appropriate features and reduce the overall number of types have to be implemented (removing types with very low frequency - among these are proper nouns or spelling errors -, removing types with very high frequency - among these are functional word like "the", "and").

As the only external knowledge base we can define linguistically motivated stopwords like articles, prepositions, conjunctions (e.g. in, this, the, of) as well as application specific stopwords (e.g. report, application, computer). All remaining items which obey some statistical restriction (frequency, selection measures) are used as features.

2.2 Offline: From Vectors to Similarities

Similar to the variety of methods for feature generation, several distance algorithms are used for the task of computing the similarity of texts. The most common methods used in work of text clustering are Cosine measure and Euclid measure.

The Cosine measure of two vectors a and b is

$$\cos(a, b) = \frac{a^T * b}{|a| * |b|} = \frac{a^T}{|a|} * \frac{b}{|b|}$$

Thus first all feature vectors are normalized to a vector of length 1 and then the scalar product is calculated between two feature vectors. Therefore the Cosine distance between two vectors remains equal even if one of the vectors is multiplied by a factor. Especially given a new text with a feature vector in which each feature has twice as large values, the distance remains equal.

One special case of this distance measure appears whenever two vectors have zero distance by pointing in the same direction in feature space.

The Euclidean distance between two vectors is

$$d_e(a, b) = \sqrt{(a - b)^T * (a - b)}$$

With this distance measure the geometrical situation in the feature space is analyzed, so that feature vectors which are far away from each other generate large distance values even if they point in the same direction.

Both measurements show different behavior when texts of different lengths are compared. In most cases for measuring text similarities the first measurement is the more adequate one. Sometimes it is better to use both distance measures and to combine the results in order to gain better results and to eliminate undesirable effects.

2.3 Offline: Combination of Different Similarity Matrices

Related work on text categorization, routing and filtering (Schütze, Hull, Pedersen 1995, Yang, Pedersen 1997, Bayer, Mogg-Schneider, Schäfer, Renz 1998, Bayer, Kressel, Mogg-Schneider, Renz 1998) has shown that a combination of different procedures gives better results than any of the procedures alone. Therefore, we also want to use a combination of different type shapes (character strings, full forms), different feature selection measures, and different weighting procedures as well as various distance algorithms for the task of restructuring texts of an intranet or text collection.

A single procedure consists of one text normalization, one type of feature generation and selection, and one distance measure. By applying one procedure, the result is a similarity matrix for the similarities of all pairs of documents. Different procedures lead to different similarity matrices. The basic idea is to analyze these different similarity matrices and generate a single one by combining the different results.

In our application we don't have a labeled data set, i.e. we don't know which similarity is the correct one for each pair of documents. Therefore it is neither possible to use statistically adapted procedures to calculate estimations for these similarities nor to optimize the combination procedure. It is only possible to use some heuristic algorithms to combine the results of the different approaches.

For example for the combination of different similarities (Similarity Combination - SC) the following calculations are sensible:

- Mean calculation: $SC = \text{average}(\text{sim}_1, \dots, \text{sim}_n)$
- Positive thinking: $SC = \text{maximum}(\text{sim}_1, \dots, \text{sim}_n)$
- Negative thinking: $SC = \text{minimum}(\text{sim}_1, \dots, \text{sim}_n)$

The basic idea for "mean" calculation is that the similarity of two documents is safer, if two or more procedures result in a high level of similarity for these documents than in case of one procedure showing a high similarity and another procedure resulting in a low similarity. The idea of "positive" is, that a similarity is valuable, if only one procedure calculates a high similarity value, while "negative" indicates that if only one procedure refuses a similarity between two documents, then we don't expect any similarity between them. The designer of a system has to decide which combination is the best for his application.

The implementation of all these combination procedures is straightforward.

We expected to get better results with mean calculation and the first short views on the results confirmed this. We combined the procedures:

- Text Normalization: case insensitive, ignoring punctuation, reducing äüö to auo
- Features1: full forms, selected by statistical means, weighted by simple tf/idf
- Distance Measure1: Cosines measure
- Features2: quad-gram (character strings of length 4), selected by statistical means, weighted by probabilistic tf/idf
- Distance Measure2: Euclidean distance
- Combination: Positive thinking

2.4 Online: Presentation

For any current document, a simple look-up in the similarity matrix gives the information if any other documents are thematically related. This online procedure, the scanning of similarities for a document, is very fast (less than 0.1 sec.).

As common to every user of intranet technology, we use browsers to open a window with links to the 'nearest' documents, i.e. those documents with highest similarity. For the comfort of the user, the 'look-and-feel' is similar to common internet search engines.

3 Experiments and Results

Up to now we adapted our system to the following three applications:

- searching for text similarities in an intranet
- clustering of customer feedback
- finding similar technical reports

Concerning the intranet of about 40,000 pages, we tried to search for additional links. This search led to very good results. For example if a user is reading a page about the company report of 1998 the system generates also a list of not linked pages of company reports of the preceding years as well as official speeches announcing these reports.

Accordingly, while reading a WWW-page about knowledge management our system automatically generates a list of related pages with topics like computer supported cooperative work or document management. These pages deal with related topics, but without explicitly containing the keyword „knowledge management“.

In the customer feedback project we got seven thousand short texts of customer feedback. After the feature generation and extraction our system was able to find clusters of interesting problems with our clustering procedure. Relating these problems to time schedule, it could be shown that some problems with the product only occurred in particular months of the year and that other problems occurred throughout all production months. Furthermore it could be shown that it is possible to group the feedback texts due to topics without any human assistance.

The third test for our system was to group automatically abstracts of technical reports of our research center. Since these reports were categorized by our documentation staff and researchers, it was possible to check the similarities of the algorithms with the pre-given categorization. Again we discovered that the combination of procedures leads to suggestions which are more reliable than every single procedure. A first assessment of our system can be based on this analysis of pre-categorized abstracts of technical reports. Here, the results are good, but we cannot conclude that this success holds in other applications as well. The only reliable measure of the system's quality is its usefulness in real applications.

4 The System at Work

The current version of the system was released to a major German company with an intranet of more than 40,000 documents. For this transfer some further monitoring and feedback processes were implemented: the monitoring makes the system robust and practical in its everyday use, the feedback gives the only real evaluation information: the opinion of 7000 intranet users.

Concerning this current version, the offline computation of the similarity matrix runs every night over a time period of approximately three hours. The presentation of thematically related documents needs less than 0.1sec. This presentation is caused by a button "What's Related?" which is part of the browser window and initiates the look-up in the similarity matrix. Fig. 3 shows the presentation of the "What's Related?" result.

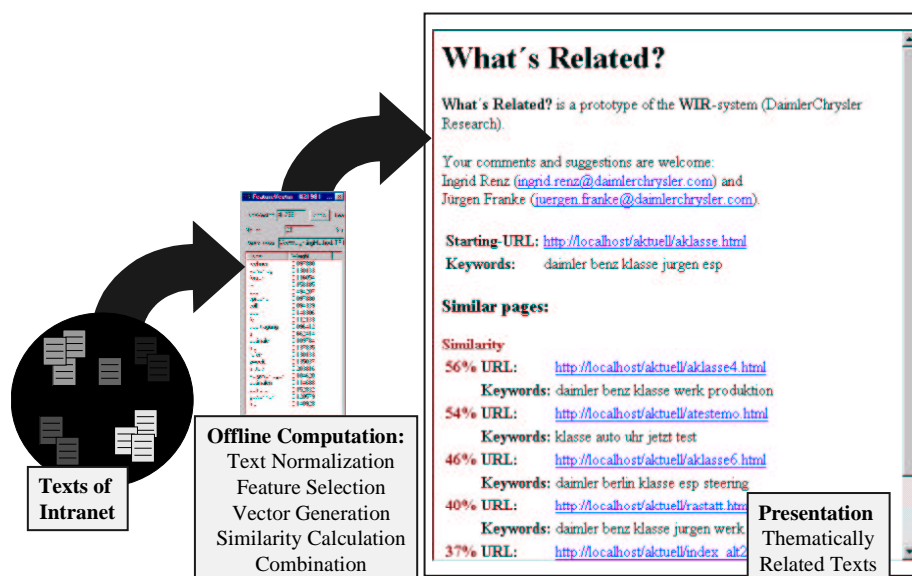


Figure 3: WIR system as knowledge management tool "What's Related?"

For the maintenance aspect of our system, changes within the intranet are recorded in a log file to check the development of the intranet. A further log file collects all searches which do not lead to a result at all. Duplicates (identical documents) as well as out-dated versions can be easily presented.

Further maintenance functions (proposal for new links, for existing categories or index terms) will be integrated later.

The evaluation of the system is a difficult matter, since the full data set is complex. Measures like precision and recall cannot be used because the information of documents that really are thematically related is not available.

For this reason, every user has the opportunity of judging the results he obtained with his query and sends his opinion as a feed-back to the system administration.

5 Conclusion and Further Directions

In this paper we gave an overview of the WIR system which at current state is fully implemented and integrated into the intranet of a large company. Although the system so far leads to good results it offers many aspects of improvement and expansion, such as the improvement of feature selection and the online computation of similarity for new documents. In this context, the system itself might overtake the initiative of proposing links to the user.

Furthermore the knowledge of text summarization techniques and document indexing might find its way into the system as well as the development of a bi- or multi-lingual system.

Based on the technology of the WIR system, various efficient and effective applications can be developed in the fields of document management, intranet applications or knowledge management, i.e. in fields where textual knowledge steadily increases and its permanent availability is crucial.

Bibliography

- Agosti, M./Crestani, F./Melucci, M. (1994). TACHIR: a Tool for the Automatic Construction of Hypertext for Information Retrieval. In *Proceedings of the RIAO 94 Conference*, New York, NY.
- Allan, J. (1995). *Automatic Hypertext Construction*. Ph.D. Dissertation, Department of Computer Science, Cornell University.
- Bayer, T./Mogg-Schneider, H./Schäfer, H./Renz, I. (1998). Daimler Benz Research: System and Experiments. Routing and Filtering. In *Proceedings of the Sixth Text Retrieval Conference (TREC - 6)*. Gaithersburg, MD.
- Bayer, T./Kressel, U./Mogg-Schneider, H./Renz, I. (1998). Categorizing Paper Documents - A Generic System for Domain and Language Independent Text Categorization. In *Journal of Computer Vision and Image Understanding*. Special Issue on Document Image Understanding and Retrieval.
- Salton, G./McGill, M. (1983). *Introduction to Modern Information Retrieval*. McGrawHill.
- Salton, G./Singhal, A./Buckley, C./Mitra, M. (1996). Automatic Text Decomposition Using Text Segments and Text Themes. In *Proceedings of Hypertext'96*, Washington, D.C. .
- Schütze, H./Hull, D./Pedersen, J.O. (1995). A comparison of classifiers and document representations for the routing problem. In *Proceedings of SIGIR 95*.
- Shin, D./Nam, S./Kim, M. (1997). Hypertext construction using statistical and semantic similarity. In *Proceeding of DL'97*, Philadelphia, PA.
- Yang, Y./Pedersen, J.O. (1997). A comparative study on feature selection in text categorization. In *Machine Learning: Proceedings of the 14th ICML*.