Textual CBR Authoring using Case Cohesion

Luc Lamontagne

Department of Computer Science and Software Engineering,
Laval University, Québec, Canada, G1K 7P4
Luc.Lamontagne@ift.ulaval.ca

Abstract. During the design phase of a textual case-based reasoning (TCBR) system, decisions have to be made regarding the internal representation of the cases and the similarity metrics. Such decisions have a significant impact on the performance of the resulting TCBR system. We believe that guidance should be provided to the designer during the authoring process. Unfortunately most of the metrics proposed in the CBR literature either require subjective human evaluation or do not apply when both problem and solution parts of the cases are textual in nature. In this paper, we propose a metric, case cohesion that can be used to provide guidance in the configuration of a case base and in the selection of appropriate similarity schemes. Case cohesion is defined from the neighborhood of textual cases and allows to estimate whether both problem and solution descriptions are well aligned. Our experimental results indicate that case cohesion can discriminate among the performances of different retrieval schemes and can also help in the unsupervised selection of the parameters required for these retrieval metrics.

1 Introduction

In the current state of case-based reasoning (CBR) research, most textual CBR systems are retrieval components. Their design usually consists of devising representation frameworks for structuring the cases ([1], [14], [15]) and selecting similarity schemes for estimating the proximity of textual cases ([2], [6]). The design of such components could benefit from the usage of performance indicators that can provide guidance to system designers during the authoring phase. These indicators could help determine whether substantial gains in terms of retrieval performance can be expected prior to the deployment of a CBR system.

Unfortunately, at the present time, the CBR domain does not provide many options for evaluating textual components. Performances of textual CBR systems are profiled using precision and recall (or a combination of both, known as F-Measure). These measures are borrowed from the information retrieval community and require subjective judgments from an external evaluator. These judgments can reveal difficult to acquire when a case base is voluminous or when no domain expert is available. It is often unreasonable to expect a human expert to validate the successive modifications made during the authoring phase of a CBR system.

Other indicators found in the CBR literature are normally used for maintenance purposes and do not transpose well to cases where both problem and solution descriptions are textual in nature.

In this paper we present some experimental results pertaining mostly to the selection of similarity schemes during the authoring phase of a CBR system (i.e. prior to deployment and validation of the system). As multiple retrieval approaches are possible, it is pertinent to determine a priori the retrieval strategy that best applies to a given case base. In [5], an approach is proposed to combine multiple similarity metrics as part of the same retrieval engine. In this work, we propose a performance indicator, case cohesion, to guide the decisions pertaining to case structuring and to the selection of a retrieval strategy. Case cohesion is based on the relatedness of the problem and solution description of cases and compares cases with respect to their similarity neighborhood.

In the next section of this paper, we briefly summarize the main design decisions for authoring textual CBR components. In section 3, we present an overview of the work pertaining to performance indicators in the CBR literature. We propose in section 4 a definition of the case cohesion indicator. We present in section 5 some experiments we conducted to assess the relevance of this indicator. We present an evaluation of three (3) retrieval configurations using case cohesion, and compare these with an a posteriori evaluation of their precision.

2 Authoring of Textual CBR Containers using Indicators

The authoring of textual CBR systems can be enhanced by using performance indicators to help make the following decisions:

- Selection of the vocabulary: What terms or features (compound terms, syntactic groups, semantic roles…) should be selected to describe the content of the cases?
- Case selection: What textual descriptions should be included in the case base?
- Case structuring: Textual cases can have various structures such as feature vectors, sequences of terms and parse trees. Since the relative importance of each term or feature can be determined using different schemes such as binary values, term frequencies, tf*idf and weight normalization, which combination should be selected?
- Selection of a similarity metrics and retrieval strategy: How to aggregate the local similarities at the feature level and to select retrieval parameters to efficiently exploit the case base?

It is important to mention that the authoring task differs significantly from case maintenance, since the latter aims to select cases in order to provide a minimal and compact case base.

In order to identify the textual CBR systems that could benefit from using performance indicators, we divide them into three categories characterized by their solution components properties:

Cases contain non textual solutions. Solutions can be identifiers, classes or numerical evaluations. Spam filtering applications [3] are representative of this category. This category offers the advantage that the mutual relevance of solutions can easily be established and the authoring of the system can be conducted based on indicators used in information retrieval and text mining.

Cases originate from complex documents and their solution is either long or ill-defined. This category relies on long textual descriptions that can not easily be compared from a gold standard point of view. Sometimes problem and solution descriptions are interleaved in the text, which makes the structuring of the cases even more complicated. Legal applications based on jurisprudence documents are examples of such systems ([2], [14]). Due to the complexity of these textual descriptions, performance indicators would not be very useful as human judgments are required to estimate the mutual relevance of the documents and to conduct an evaluation of the resulting CBR systems.

Cases contain short textual solutions. Problem and solutions are distinct descriptions that contain few sentences. Sometimes, the vocabulary used to describe these cases is limited and facilitates the estimation of their relevance. Frequently asked questions [2], email response [6] and some incident reports (for example [12]) have such a structure. Due to the limited complexity of the textual descriptions, this category can potentially benefit from guidance during the authoring phase. This is the issue we explore in the rest of this paper.

3 Related work in the CBR literature

The work presented in this paper relies on the use of indicators to guide the construction of a TCBR system. We have identified various metrics from the CBR literature to characterize a case base and Table I contains a list of the most prominent ones. Most of them were proposed for structural CBR systems and are used for the maintenance of a case base to determine which cases should be retained or removed from a case base.

Furthermore, the definition of these measures is not well adapted to Textual CBR: Many indicators require the case to be structural, to be homogeneous (defined from a limited number of attributes) or to have a limited number of possible values for each of the attributes. However, a textual case is by definition heterogeneous, its internal representation contains few terms with respect to the full vocabulary of the system, and the overlap between cases is low because the descriptions are seldom repetitive.

Moreover, the measures related to case adaptation (coverage and reachability) are impractical in the current state of the domain. Adaptation in Textual CBR has received little attention and these measures cannot be defined easily for cases relying on textual descriptions.

Table 1. Some performance indicators used the CBR literature

Indicators	Refs	Comments
Precision, recall and F-measure	[4]	Used to characterize the efficiency of a retrieval module, i.e. the capacity to retrieve mostly and solely relevant cases. The application of these indicators is limited by the availability of human judgments when solutions are textual in nature.
Case Density	[10]	Indicates the average proximity of cases in a case base. It also estimates the competence of a case base to solve problems as a higher case density involves a greater concentration of cases in a limited problem space. Hence, the contribution of each individual case is more restricted. This can be estimated from case similarity.
Case Distribution	[13]	This is the distribution of the cases over a set of possible problems. An irregular distribution indicates that some problems may not be solved. This indicator is difficult to apply in a textual setting.
Uniqueness and Redundancy	[8], [9]	Uniqueness determines whether no other case contains the same problem and solution descriptions. Case redundancy is either a) the exact matching of cases, b) a subsumption of one case by another, or c) a high similarity between cases. This is useful for case maintenance to determine the cases to remove in order to improve the performance of a system.
Consistency	[8], [9]	Cases having the same problem descriptions should not have different solutions. Definitions of consistency are usually based on a) the coverage between descriptions and b) some domain rules to define the combination of values or attributes deemed conflicting by the system designer.
Coverage and reachability	[11]	This is defined for CBR systems with adaptation features. It designates the extent of problems that can be solved by a system. These measures combine the capacity to determine the nearest neighbors of a target case and to evaluate whether neighbors can be adapted to reconstruct the target solution. Not applicable to current TCBR systems.
Regularity	[7]	Regularity is proposed to describe the relationship between problem and solution descriptions to ensure that similar problems have similar solutions.

4 Case Cohesion Indicator

In order to conduct our experiment regarding the selection of an appropriate retrieval scheme, we propose an indicator called *cohesion(c)*. The cohesion indicator is intended to measure the degree of relatedness of problem and solution descriptions of a textual case. To define this indicator, we assume that a case presents a strong cohesion if, in a given neighborhood, some other cases present similar problem and solution relationships.

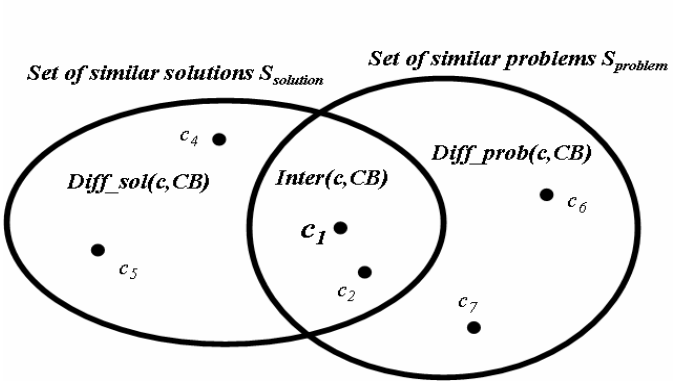

Fig. 1. Sets of cases having similar problems and/or solutions.

To be more specific, let us consider two sets related to a specific case c_1 as depicted in Figure 1: The set of cases having similar problems ($S_{problem}$) and the set of cases having similar solutions ($S_{solution}$). These two sets can be defined using similarity thresholds δ_{prob} and δ_{sol} :

$$S_{problem}(c_1, CB) = \{ c \in CB: sim_{prob}(c_1^{prob}, c^{prob}) > \delta_{prob} \} \quad (1)$$

$$S_{solution}(c_1, CB) = \{ c \in CB: sim_{sol}(c_1^{sol}, c^{sol}) > \delta_{sol} \} \quad (2)$$

where sim_{prob} and sim_{sol} are respectively the similarity of the problems and of the solutions of two cases. In our textual setting, the measures used to compute these similarities are those considered by the system designer during the authoring of the CBR system. We will revisit this issue in the next section.

From both sets, we can determine three regions with cases having:

- both solutions and problems similar to c_1

$$\text{Inter}(c_1, CB) = S_{problem}(c_1, CB) \cup S_{solution}(c_1, CB) \quad (3)$$

- only problems similar to c_1

$$\text{Diff\_prob}(c_1, CB) = S_{problem}(c_1, CB) - \text{Inter}(c_1, CB) \quad (4)$$

- only solutions similar to c_1

$$\text{Diff\_sol}(c_1, CB) = S_{solution}(c_1, CB) - \text{Inter}(c_1, CB) \quad (5)$$

These three sets are then used to measure the quality of the relationship between a case and its neighborhood of problems and solutions. The union corresponds to the number of distinct cases contained in the three sets.

The degree of case cohesion can then be defined as follows:

$$\text{cohesion}(c_1) = \text{Inter}(c_1, CB) / \text{Union}(c_1, CB) \quad (6)$$

Hence, a case c_1 has strong cohesion if its behavior is similar to those of the cases in its neighborhood. This measure indicates whether the relationship between a problem and its solution bears resemblance to the relationships found within the other cases. We propose to use this definition to measure the extent to which textual problems and solutions are well aligned. We believe that a larger proportion of well aligned cases is indicative of a good case structure and an adequate similarity measure. This is not true of weakly cohesive cases that may reveal to be unique, specific, inconsistent, incoherent or ill-structured, resulting in additional authoring efforts. We do not address this issue in this paper as our goal is to determine whether case cohesion reveals a good indicator for guiding component selections during the authoring phase.

5. Experiments and Evaluation Results

5.1 The Textual CBR Configurations Compared

As part of our experiments, we compare three (3) textual CBR configurations we studied in previous work to determine whether case cohesion results corroborate performance evaluations in terms of precision. These configurations are:

*Configuration A - Cosine, vector normalization & TF*IDF weights:* This is a configuration used in information retrieval and is frequently encountered in textual CBR. To obtain such a configuration, we first determine the vocabulary of the CBR system as a subset of the words contained in each case description. In this paper, we present results for a vocabulary containing words that appear at least 3 times in the corpus. We then obtain case problem and solution descriptions by converting the corresponding document to a term vector. Weights are then assigned to each feature according to a tf*idf function, which is based on the frequency and the distribution of the words in the documents. Finally term vectors are normalized to ensure that similarity values are within a [0,1] range when computed using a cosine metric. Additional details on these techniques can be found in [4].

Configuration B- Cosine & Term vector normalization: This is similar to the previous configuration where the tf*idf weights are replaced by term frequencies. This is used to compare configurations that differ only in the internal representation of the case base and not in the similarity schemes.

Configuration C - Case expansion: This retrieval approach was proposed in [6] and can be summarized by the following:

- Determine terms that co-occur in the documents used to build the cases. Cooccurrence indicates that the presence of one word influences posi-

tively the presence of another word. Mutual information is used to estimate the statistical significance of the word cooccurrences.

- For each word of the vocabulary, build cooccurrence lists that are used to expand term vectors. Lists only contain words with mutual information superior to some threshold arbitrarily selected by the designer of the CBR system. Hence this parametric approach requires some tuning during the authoring phase.
- Structure the case base as in Configuration A and expand cases by adding terms from the cooccurrence lists of each word contained in the descriptions.
- Compute similarity using a cosine function as for Configuration A.

5.2 Procedure for Comparing the Configurations

As described in Section 4, case cohesion requires two thresholds (δ_{prob} and δ_{sol}) to determine the problem and solution neighborhoods of each case ($S_{problem}$ and $S_{solution}$). To determine how each configuration behaves with respect to case cohesion, we vary these thresholds and compare the performance profiles obtained for each configuration. In our experiment, we have followed the following approach:

a) We use each case as a target case to find the set of solutions in the case base with a similarity value exceeding δ_{sol};
b) Given this set of solution, we vary the similarity threshold δ_{prob} of the problems to identify the set of problems that provides a maximum cohesion (i.e. find max $_{\delta prob}$ $cohesion(c_{target})$);
c) By repeating this experiment for different target cases and different solution thresholds, we obtain curves of maximal cohesion for each configuration;
d) We visually compare the curves of maximum cohesion for each configuration in order to determine whether a trend can be identified.

As an example, we present in Figure 2 cohesion profiles obtained for different values of threshold δ_{prob} and δ_{sol} for the Case Expansion configuration. Maximal cohesion is obtained by taking the maximal value for each of the problem thresholds. The mutual information value used for pruning the cooccurrence lists was set to 0.0 for producing the result presented in this figure.

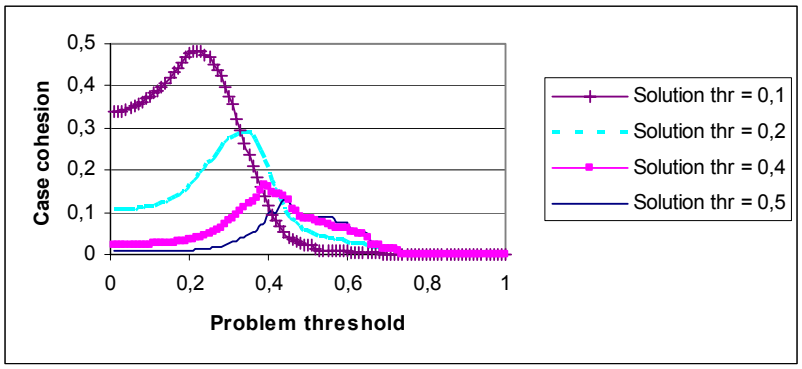

Fig. 2. Examples of cohesion profiles for the Case expansion configuration

5.3 Evaluation Results

We obtained our experimental results using a set of 103 cases taken from an email response application we developed. A term vector representation and a cosine measure, as described in configuration A, were used to structure solutions, to estimate their similarities, and to obtain the set of similar solutions $S_{solution}$ for each case. Hence, to conduct our experiments, solution similarity was defined at the lexical level.

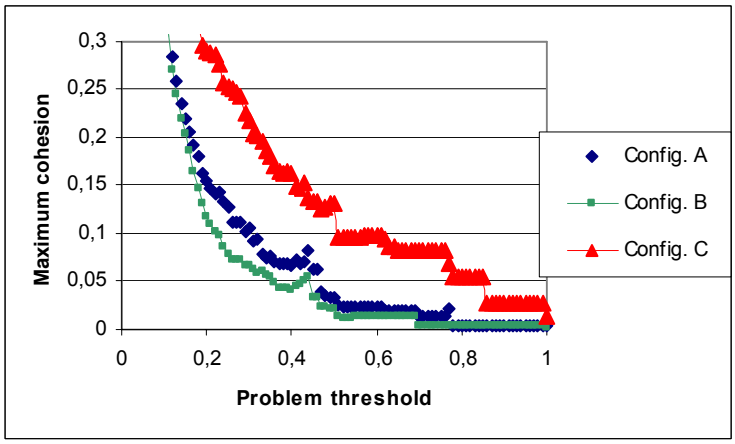

Fig. 3. Maximum case cohesion profiles for the three configurations.

In figure 3, we present maximal case cohesion results obtained for the three proposed configurations. We first note that case expansion (Configuration C) obtains cohesion values superior to those of a combination of tf*idf weights & cosine metric (Configuration A). This observation holds for each of the threshold values. Hence, the cohesion metric suggests the selection of configuration C whatever conditions the CBR system is to be operated in. This results support evaluation results obtained in a previous study [6] which concluded that the case expansion approach reveals to be a better choice than a simple tf*idf configuration of our case base. In this study, evaluations based on human judgments resulted in a precision of 63% for Configuration C and 57% for Configuration A.

We also note that replacing tf*idf weights (Configuration A) by term frequencies (Configuration B) brings a slight degradation in case cohesion. This also concurs with a posteriori evaluation conducted on the precision of both configurations (Configuration A – 57%, Configuration B – 54%).

5.4 Threshold Selection for Case Expansion

As seen in the description of the configurations, the case expansion configuration is parametric since it requires a threshold on mutual information (MI_thr) to truncate the cooccurrence lists. We verify in the section whether case cohesion can provide guidance in the selection of this parameter.

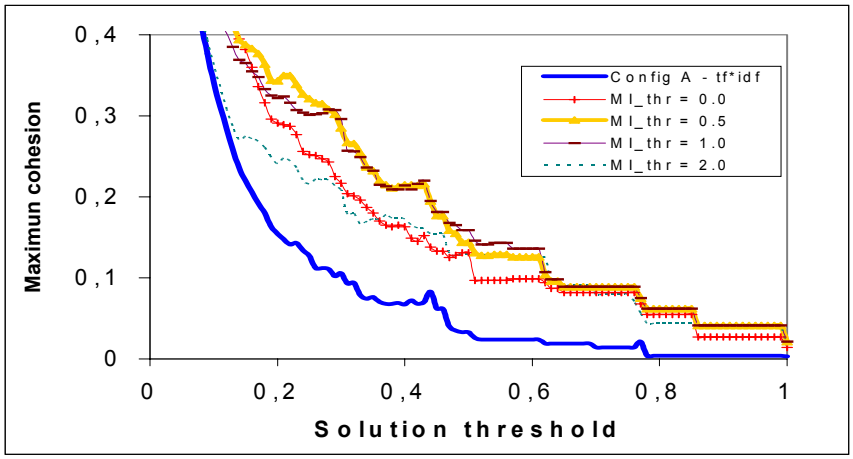

Fig. 4. Maximum cohesion corresponding to various mutual information thresholds.

By reproducing the same type of analysis as the one conducted in the previous section, and by varying the mutual information threshold (MI_thr), we obtain the maximum cohesion profiles presented in Figure 4.

It clearly appears from these results that the expansion approach gives superior results to configuration A, independently of the MI threshold value used to expand cases. It is also interesting to note that mutual information thresholds between 0.5 and 1.0 provide maximal results. These two observations are again in agreement with an evaluation of precision of this retrieval configuration presented in Figure 5. Results for Configuration C are presented for various runs with different minimal term frequencies to insert a word in the vocabulary.

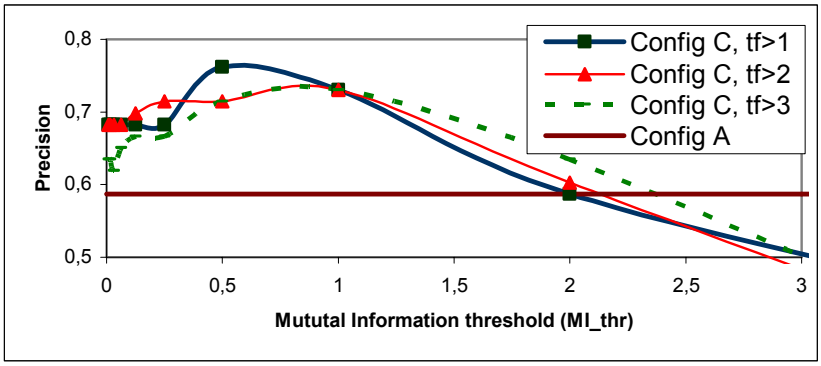

Fig. 5. Precision of Configurations A and C with different minimal term frequencies (tf).

6 Conclusion

In this paper, we conducted some experiments to determine whether it is possible, without human judgment pertaining to the mutual relevance of cases, to make decisions regarding the structure of a case base and the selection of a retrieval scheme. These decisions are typically taken during the authoring phase of a CBR system. We proposed a simple performance indicator called case cohesion defined from the similarity neighborhood of the cases. As a result of our experiments, case cohesion can be used to discriminate among various retrieval schemes since the trends identified in the cohesion profiles seem to corrobate system precision based on human judgments.

Our results are preliminary and the influence of various factors should be further investigated in future work. First, the similarity between solutions was estimated using a tf*idf and cosine combination. We do not know if case cohesion is sensitive to the similarity measure used to determine the set of similar solutions. Also our case descriptions are rather short (a few sentences) and rely on a limited vocabulary. What would happen if the textual descriptions were more complex? Finally, we should investigate how other indicators could be defined from the set of similar problems and solutions to provide guidance on other aspects of the authoring process.

References

1. Brüninghaus, S.; Ashley, K. D. (1999); "Bootstrapping Case Base Development with Annotated Case Summaries", in *Proceedings of the third International Conference on Case-based Reasoning ICCBR-99*, LNAI 1650, pp. 59-73.
2. Burke R., Hammond K., Kulyukin V., Lytinen S., Tomuro N., Schoenberg S. (1997) "Question Answering from Frequently-Asked Question Files: Experiences with the FAQ Finder System", *AI Magazine*, 18 (2), pp. 57-66.
3. Delany; S.J.; Cunningham, P.; Coyle, L. (2005) "An Assessment of Case-Based Reasoning for Spam Filtering", *Artificial Intelligence Review Journal* , 24(3-4) pp. 359-378.
4. Jurafsky, D.; Martin, J. H. (2000); *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics*. Prentice-Hall.
5. Lamontagne, L.; Abi-Zeid, I. (2006) "Combining Multiple Similarity Metrics Using A Multicriteria Approach", to appear in *Proceedings of ECCBR 2006*, LNAI 4106, Springer-Verlag.
6. Lamontagne, L.; Langlais, P.; Lapalme, G. ; (2003) "Using Statistical Models for the Retrieval of Fully-Textual Cases", in Russell, I.; Haller, S. (Editors), *Proceedings of FLAIRS '03*, AAAI Press, Ste-Augustine, Florida, pp.124-128.
7. Leake D. B.; Wilson, D. C. (1999) "When Experience is Wrong", in *Proceedings of the third International Conference on Case-based Reasoning ICCBR-99*, LNAI 1650, Springer-Verlag, p. 203-217.
8. Racine K.; Yang Q. (1997) "Maintaining Unstructured Case Bases", in *Proceedings of the Second International Conference on Case-Based Reasoning ICCBR-97*, LNAI 1266, Springer-Verlag, pp. 553-564.
9. Reinartz, T.; Iglezakis, I.; Roth-Berghofer, T. (2000); "On Quality Measures for Case Base Maintenance", in *Proceedings of the 5th European Workshop on Case-Based Reasoning*, LNAI 1898, Springer-Verlag, pp. 247-259.
10. Smyth, B.; McKenna, E. (1998) "A portrait of case competence: Modelling the competence of case-based reasoning systems". In *Proceedings of the Fourth European Workshop on Case-Based Reasoning*, Springer, pp. 208-220.
11. Smyth, B.; McKenna, E. (1999) "Building Compact Competent Case-bases", in *Proceedings of the third International Conference on Case-based Reasoning ICCBR-99*, LNAI 1650, Springer-Verlag, pp. 329-342.
12. Varma, A. (2001) "Managing Diagnostic Knowledge in Text Cases", in *Proceedings of ICCBR'2001*, LNAI 2080, Berlin: Springer, pp. 622–633.
13. Watson, I. (1997) *Applying Case-Based Reasoning: Techniques for Enterprise Systems*, Morgan Kaufmann Publishers Inc.
14. Weber R.; Martins A.; Barcia R (1998) "On legal texts and cases", *Textual Case-Based Reasoning: Papers from the AAAI-98 Workshop*, Technical Report WS-98-12, AAAI Press, pp.40-50.
15. Wiratunga, N., Koychev, I., Massie, S. (2004), "Feature Selection and Generalisation for Retrieval of Textual Cases". in *Proceedings of the 7th European Conference on Case-based Reasoning*, Springer-Verlag, LNAI 3155, pp 806--820.