Enhancing the Performance of Feature Selection using a Hybrid Genetic Algorithm

N. Vanjulavalli, PhD
Assistant Professor,
Department of Computer Science,
Annai College of Arts and Science,
Kumbakonam.

A. Kovalan, PhD
Assistant Profeessor (S.S),
Department of Comp. Sci.and Applications
Periyar Maniammai University,
Thanjavur.

ABSTRACT

Information Retrieval (IR) issues have attracted increasing attention due to the growing availability of the documents. The retrieval of web pages is more challenging due to the ambiguous nature of the unstructured information found in these pages. Ontologies help to overcome the disambiguate nature of the natural language by the use of standard terms that relate to specific concepts. Ontology is a hierarchy of concepts with attributes and relations that defines an agreed terminology to describe semantic networks of interrelated information units. Ontology provides a vocabulary of classes and properties to describe a domain, emphasizing the sharing of knowledge and the consensus about its representation. This research focuses on IR systems moving from a lexical to semantic interpretation to match object and queries on a semantic basis. In natural language, many words are ambiguous giving different meanings based on the context and situation. Therefore, development of web directories, classification of web pages and analysis of topic-specific search are useful. Classification of contents makes an important part of most of the content management and retrieval activities. The underlying objective of this research work is to develop an effective and efficient feature selection and classification algorithm that can achieve good accuracy in classifying web pages.

Keywords

Information Retrieval, Feature Selection, Genetic Algorithm

1. INTRODUCTION

Information Retrieval (IR) concerns science and technology and effective retrieval of data by interested parties from an information repository. The problem in IR is the quest to locate information resources among large repositories, satisfying information need which is expressed through a query by the user. Information resources are represented as objects (items) in a medium such as text, image, audio, or combination of all three.

Figure 1.1 shows the basic steps of IR process. In the conceptualization step of IR, user represents desired information using query syntax. Information complexity needs only partially reflected in query. In reality, user approximates query need by representing several for outstanding need characteristics so that the query is represented as a terms set related to some coordinating terms/symbols. This usually causes ambiguities in retrieval; so it serves as main source for refinement process.

In the retrieval step of IR, query is implemented against underlying information repository using retrieval model like Boolean, vector space or probabilistic model [7]

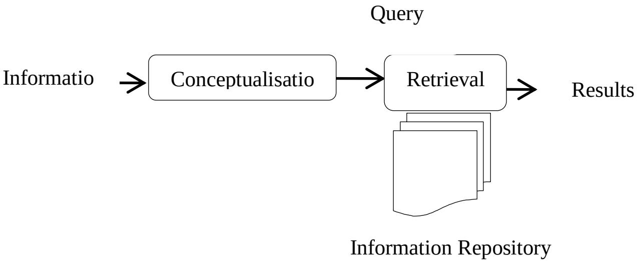


Figure 1.1 Basic IR Process

Features extracted are the key for achieving good classification of documents in IR. Feature selection is a problem to be addressed in artificial intelligence where the issue is in developing feature selection techniques so as to choose a small feature set to reduce a system's cost and running time, and to achieve acceptably high recognition rate. This led to development of various techniques to select an optimal features subset from a larger, possible features set. Such techniques are classified into two categories as follows:

Problem specific strategies are developed in the first approach based on domain knowledge to reduce features number used in a manageable size. The second approach is used when domain knowledge is unavailable or when exploiting cost is high. Here, generic heuristics, effective greedy algorithms select a subset "d" of available "m" features [11].

Feature selection is also called attribute selection. Here the need is to locate an optimal feature subset usually not easily controlled or directed. Many problems related to feature selection are NP-hard. Basically, this process consists of four different steps as seen in figure 1.3, including subset generation, subset evaluation, stopping criterion and result validation

2. ONTOLOGY

Ontology is a specification of an abstract which represents a simplified view for a purpose. Ontology is defined as a set of representational terms called concepts. Concepts interrelationship describes a target world. Ontology is built in two ways, domain dependent and generic. Some generic ontology examples are Cyc, Word Net, and Sensus [10].

Ontology is a hierarchy of concepts with attributes and relations that defines an agreed terminology to describe semantic networks of interrelated information units. Ontology provides a vocabulary of classes and properties to describe a domain, emphasizing the sharing of knowledge and the consensus about its representation. For instance, ontology about *Computer applications* could include classes such as *Software*, *Document*, *Person*, and properties (relations) like Person *creator* of a document, software *depends on* software, or software *generates* document. The goal is then to describe services and contents by a network of nodes typified and interconnected through classes and properties defined in shared ontologies. Thus, for example, once ontology about computer applications had been created, a virtual company could organize its contents defining instances of applications, developers, documents, etc. A software agent bro sing a network like that might recognize the different information units, obtain specific data or reason about complex relations [12].

Ontology refers to the structured representation of the domain knowledge which includes defining of classes, relations and functions among the objects [13]. Ontology models the relationship between the concepts and objects for a domain. IR for semi structured data such as web pages is challenging due to the ambiguous nature of the unstructured information found in these pages. During IR, words in natural language may have different meanings depending on the context leading to inefficient retrieval [14]. In ontology, the context of vocabulary is represented and constrained in the ontology model, thus, overcoming the disambiguous meanings of words in the free text.

2.1 Ontology learning

Ontology learning refers to extracting ontological elements like conceptual knowledge from input and constructing ontology from it. Ontology learning aims at semi-automatically or automatically building ontology's from a text with limited human exertion. It is also a set of methods/techniques used to build ontology from scratch, enriching/adapting an existing ontology semi-automatically using many sources. It uses methods from diverse fields like knowledge acquisition, machine learning, IR, natural-language processing, artificial intelligence, reasoning and database management.

2.2 Ontologies used for Query Expansion

Query expansion is the method of supplementing the user's query with additional terms to improve results during retrieval. Similar and pertinent terms to query terms are usually used for expansion. Two common strategies used to find expansion terms are adding related terms based on relatedness measure and based on relevance feedback.

The two main approaches to query expansion are probabilistic query expansion and ontology query expansion. Probabilistic query expansion more widely used and is based on calculating co-occurrences of terms in documents and selecting terms that are most related to query terms. Ontological methods use semantic relations drawn from the ontology to select terms. The following describes how different ontologies are used for query expansion.

Query Expansion with General/Domain-specific Ontologies

Query Expansion with Spatio-temporal Ontology SAPO [15].

3. RELATED WORKS

1. For years people realized the importance of archiving and locating information. Computers made it possible to store large amounts of information; and locating useful information from such collections became necessary. The field of IR was born out of this necessity in the 1950s. In the last forty years, it matured considerably. Many IR systems are used daily by various users. A brief overview of key advances in IR and a description of where state-of-the-art was in the field were presented by [16].

An overview and instruction on evaluation of interactive IR systems with users was presented by [2] the aim of which was to catalogue related material into one source. The article reveals historical background on development of user-centered approaches in evaluating interactive IR systems; describes components of interactive IR system evaluation; shows various experimental designs and sampling strategies; presents core instruments, measures data collection techniques; explains data analysis techniques and reviews/discusses earlier studies. This article discussed validity/reliability issues regarding measures and methods, presented research ethics background information and discusses ethical issues specific to interactive IR studies. Finally it ends discussing outstanding challenges and future research.

2. Static index pruning methods that reduced index size in IR systems was introduced by [1]. Investigation of uniform and term-based methods that removed selected entries from index had only minor effect on retrieval. There was a fixed cut-off threshold in uniform pruning and index entries contribution to relevance scores bounded by a threshold was removed from index. Cut-off threshold was determined for each term in term-based pruning and varied with each from. Experimental evidence existed for each compression level and term-based pruning outperformed uniform pruning under various precision measures. Final presentation was theoretical/experimental evidence that under term-based pruning it was possible to prune index greatly and still get retrieval almost as good as that based on full index.

3.Statistical language modeling was successfully used in speech recognition, part-of-speech tagging and syntactic parsing and more recently to IR. According to the new paradigm, each document is viewed as a language sample, and queries a generation process. Retrieved documents were ranked based on probability of producing a query from corresponding documents language models. [17] presented a new language model for IR based on data smoothing techniques range including Good-Turing estimate, curve-fitting functions, and model combinations. The conceptually simple and intuitive model could be incorporate probabilities of phrases like word pairs and word triples. Experiments with Wall Street Journal and TREC4 data sets revealed that the new model's performance as comparable to that of INQUERY and improved than that of another IR language model. Specifically, word pairs improved retrieval performance.

Treebanking Decisions Feature (TDF) is based on a candidate trees set created by language grammar and disambiguation is by annotators . This reduces man power to create tree and better annotation is built. TDF improves annotation by humans and evaluates differences between individual's analyses. After creating n number of trees, sentences are applied to trees to decide about ambiguities. Usually n candidate trees produce approximately log (n) decisions for a sentence. Decisions are similar to accurate judgments by human annotators who created decision trees.

4.FEATURE SELECTION

Any classification technique's performance depends on features of training and test data sets. Feature selection also called variable selection, feature reduction, attribute selection or variable subset selection, is a common machine learning technique to select a relevant features subset to build robust learning models. In machine learning approaches, feature selection is an optimization issue involving selection of an appropriate feature subset.

Generally, feature selection is formulated under single objective optimization framework and stated as follows: In a set of features S and classification quality measure P, determine feature subset F^* so that:

$$P\left(F^*\right) = max_{F \in S} P\left(F\right) \qquad (4.1)$$

Generally, search space for such problems is 2^d, where d is total number of possible features. So, exhaustive search strategies are inappropriate in this case. Heuristics based techniques like GA are used to search for appropriate feature combination [18].

Feature subset selection identifies and removes irrelevant and redundant information. This reduces data dimensionality and allows learning algorithms to operate quicker and better. In some cases, feature classification accuracy is improved; in others, the result is compact, easily interpreted representation of target concept.

Feature selection algorithms (with notable exceptions) perform search through features space, and so must address 4 issues affecting search nature (Langley 1994):

1. Starting point. The point from which to begin search in the feature subset space which can affect search direction. An option is beginning with no features and successively adding attributes. Here, search proceeds forward through search space. Conversely, search begins with features and successively removing them. Here, search proceeds backward through search space. An alternative is beginning in the middle and moving outward from that point.
2. Search organization. An exhaustive feature subspace search is prohibitive for all but an initial features number. With N initial features, there exist 2^N possible subsets. Heuristic search strategies are feasible than exhaustive ones and ensure good results, though not guaranteeing location of optimal subset.
3. Evaluation strategy. How feature subsets are evaluated is the biggest differentiating factor among feature selection machine learning algorithms. A paradigm, dubbed *filter [19]*; operates independent of a learning algorithm—undesirable features are filtered out from data prior to learning. Such algorithms use general data characteristics based heuristics to evaluate feature subsets merit. One school of thought argues that a specific induction algorithm's bias should be considered when selecting features. Called the *wrapper [19]*, uses induction algorithm with statistical re-sampling technique like cross-validation to estimate final feature subsets accuracy.
4. Stopping criterion. A feature selector decides when to stop searching through feature subsets space. Depending on evaluation strategy, feature selector may stop adding/removing features when no alternative improves a current feature subset's merit. Alternatively, the algorithm might revise feature subset till merit does not degrade. Another option is continuing generating feature subsets till reaching opposite search space end and then selecting the best.

5. METHODOLOGY

In the proposed features extraction, the features are extracted based on the ontology and feature selection is achieved by GA. A concept based tree structure is built on a generalisation/specialisation relationship to conceptualization the domain. Browsing knowledge is made easier if the conceptual architecture of the knowledge based is identified as a whole and information is accessible by intra conceptual hierarchical links during browsing. Thus, when browsing in a vast information base, data mapping provides interesting solutions in representing the data [20]. This is also applicable to semantically annotated knowledge bases resulting in concepts tree structure. The concepts are organized into a taxonomy tree where each node represents a concept and every concept a specialization of its parent.

Mutation aims to introduce new genetic material in existing chromosomes. It also occurs at a probability pm, called mutation rate. A small value for pm _ (0, 1) ensures good solutions are not distorted much. Conventional mutation operator is performed on gene-by-gene basis. With a given mutation probability, every gene in chromosomes in a population undergoes mutation. To further prevent much distortion of solutions, the mutation operator is modified by allowing only the genes for mutation that satisfy the condition "((*K* <*Kmax*) and (its allele = '0')) or ((*K* >*Kmin*) and (its allele = '1'))", where *K* is the number of clusters (or 1's) for the chromosome being examined for mutation. If a gene is selected for mutation then its allele is altered. The term "altering the allele" means to change an allele '0' to '1' or vice versa. The modified version of mutation operation is illustrated as follows:

```
// Mutation under given conditions:
for$i$= 1 to $P$ do
for each gene $g$ of $Chi$
If (($Ki$ <$Kmax$) and (allele of $g$ = '0')) or (($Ki$ >$Kmin$) and (allele of $g$ =    '1')) then
Generate a random (float) number γ from the range (0, 1).

If γ < pm then mutate the allele of $g$.
```

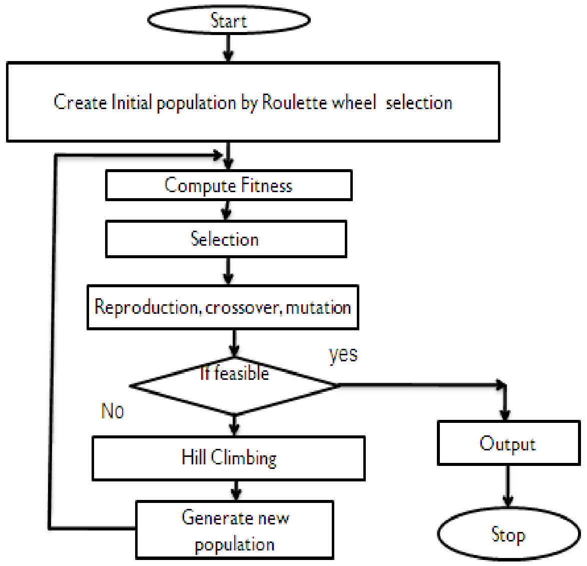
5.1 Fitness function

Feature subset selection's goal is to use fewer features to achieve same or better performance. Hence, fitness evaluation has two terms: (i) accuracy and (ii) features number are used. Classifier's performance is estimated using validation data set to guide GA. Each feature subset contains certain features. When two subsets achieve same performance having different features numbers, a fewer features subset is preferred. Between accuracy and feature subset size, the former is a major concern.

5.2 Proposed GA-Hill Climbing Optimization

The GA tends to get trapped in the local minima, Thus to overcome this problem, Hill Climbing is used as local search in the hybrid algorithm. The hybrid optimization of GA and the Hill Climbing algorithm starts with generating of initial populations of GA. The GA process such as selection, crossover and mutation is performed. Finally the best individuals are selected and saved. This process gets repeated till it reaches the stopping criteria of GA. Once the stopping condition of GA process is met, then the Hill Climbing process gets started.

Hill climbing optimization has 4 input parameters like, objective function, starting points, range and step of the search. Search space for hill climbing is spanned by transformation parameter basis. Search space basis is usually an orthogonal set or non-degenerated [22]. Rigid body rotation is orthogonal. Rotation and translation are correlated, as rotation around an arbitrary point can decompose into rotation around origin plus a translation. Affine transformation is not orthogonal, but is non-degenerated.



Flow Chart of Proposed Methodology

6. RESULTS AND DISCUSSION

The proposed genetic based feature extraction for web page classification is assessed using the 4 Universities Dataset and compared with IDF feature extraction method. Classification accuracy, Recall and precision are measured for both proposed and IDF techniques. In this study, the Bagging is done with REPtree, BFtree, J48, and CART.The accuracy, precision, recall and f measure are computed as follows:

$$Accuracy\ (\%) = (TN + TP)\ /\ (TN + FN + FP + TP) \quad (4.7)$$

$$precision = \frac{TP}{TP + FN} \quad (4.8)$$

$$recall = \frac{TP}{TP + FP} \quad (4.9)$$

$$f\ \ Measure = \frac{2 * recall * precision}{recall + precision} \quad (4.10)$$

where TN (True Negative) = Number of correct predictions that instance is invalid

FP (False Positive) = Number of incorrect predictions that instance is valid

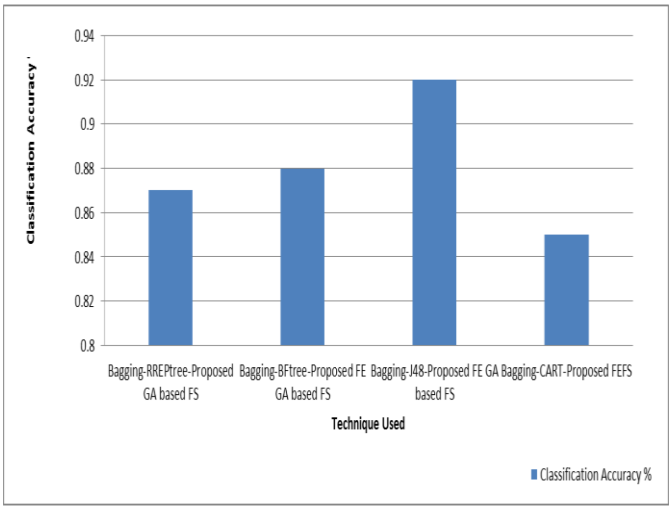
FN (False Negative) = Number of incorrect predictions that instance is invalid

TP (True Positive) = Number of correct predictions that instance is valid

Table 4.1 Classification Accuracy and Root Mean Squared Error

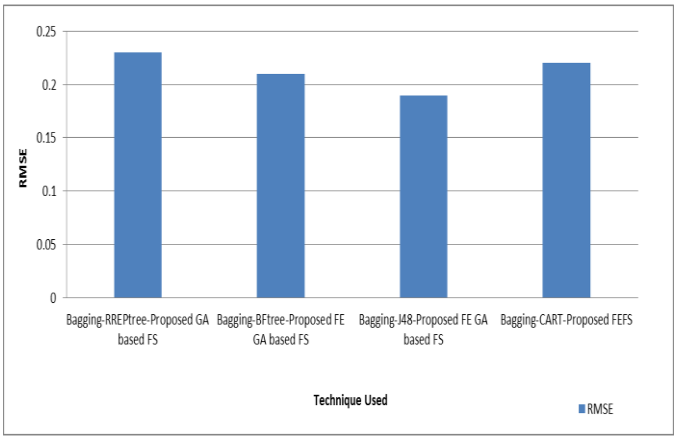
Method Used	Classification Accuracy	RMSE
Bagging-RREPtree-Proposed GA based FS	0.87	0.23
Bagging-BFtree-Proposed FE GA based FS	0.88	0.21
Bagging-J48-Proposed FE GA based FS	0.92	0.19
Bagging-CART-Proposed FEFS	0.85	0.22

Classification Accuracy

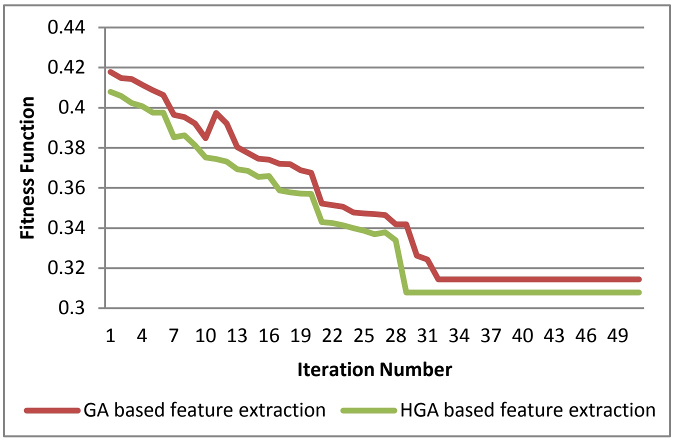


RMSE

Convergence occurred at iteration number 155 for GA based Feature Extraction and Convergence occurred at iteration number 140 for HGA based Feature Extraction.



Fitness Function



7. CONCLUSION

Ontology-Based Information Extraction is a widely researched information extraction sub field. In this paper, ontologies are used for information extraction process. Features are extracted using IR approaches such as IDF and proposed ontology based features. The extracted features are processed using GA to find optimal feature subset which is used as the input for the classifiers. In order for the GA to select a subset of features, a fitness function must be defined to evaluate the performance of each subset of features. GA explores the space of subset of features to try to find a minimum subset of features with good classification performance.

The feature subset is classified using bagging with various decision trees (REPtree, BFtree, J48, and CART). The experimental results show that proposed feature extraction improves the precision and recall satisfactorily. The Hybrid GA based feature selection achieves better classification accuracy ranges from 0.27% to 1.7% than GA based feature selection.

8. REFERENCES

[1] Carmel, D., Cohen, D., Fagin, R., Farchi, E., Herscovici, M., Maarek, Y. S., and Soffer, A. Static index pruning for information retrieval systems. *In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval ACM* ,2001, pp. 43-50.

[2] Cheng, C. K., Pan, X., andKurfess, F. (2004). Ontology-based semantic classification of unstructured documents. In Adaptive Multimedia Retrieval (pp. 120-131). Springer Berlin Heidelberg.

[3] Doan, A., Madhavan, J., Domingos, P., and Halevy, A. (2004). Ontology matching: A machine learning approach. Handbook on Ontologies in Information Systems, 397-416.

[4] Ekbal, A., Saha, S., andGarbe, C. S. (2010, August). Feature Selection Using Multiobjective Optimization for Named Entity Recognition. In Pattern Recognition (ICPR), 2010 20th International Conference on (pp. 1937-1940). IEEE.

[5] Fernandez, M., Cantador, I., López, V., Vallet, D., Castells, P., and Motta, E. (2011). Semantically enhanced Information Retrieval: an ontology-based approach. Web semantics: Science, services and agents on the world wide web, 9(4), 434-452.

[6] Kelly, D. Methods for evaluating interactive information retrieval systems with users, *Foundations and Trends in Information Retrieval*, 3(1—2), 2009, pp.1-224.

[7] Khan, L., McLeod, D., andHovy, E. (2004) " Retrieval effectiveness of an ontology- based model for information selection", The VLDB Journal—The International Journal onVery Large Data Bases, 13(1), 71-85.

[8] Kohavi, R. (1995). Wrappers for performance enhancement and oblivious decision graph(Doctoral dissertation, Stanford university).

[9] Pan, X., andAssal, H. (2003, October). Providing context for free text interpretation. In Natural Language Processing and Knowledge Engineering, 2003. Proceedings. 2003 International Conference on (pp. 704-709). IEEE

[10] Shen, D., Chen, Z., Yang, Q., Zeng, H. J., Zhang, B., Lu, Y., and Ma, W. Y, (July 2004) "Web-page classification through summarization" In Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 242-249). ACM,.

[11] Shibu, S., Vishwakarma, A., and Bhargava, N, combination approach for Web Page Classification using Page Rank and Feature Selection Technique. *International Journal of Computer Theory and Engineering*, 2(6), pp.897-900, 2010.

[12] Singhal, A. (2001). Modern information retrieval: A brief overview. IEEE Data Eng. Bull., 24(4), 35-43.

[13] Song, F., and Croft, W. B. (1999, November). A general language model for information retrieval. In Proceedings of the eighth international conference on Information and knowledge management (pp. 316-321). ACM.

[14] Song, F., and Croft, W. B. A general language model for information retrieval. *In Proceedings of the eighth ACM. international conference on Information and knowledge management* Nov. 1999, pp. 316-321.

[15] Song, L., Mi, H., Lü, Y., and Liu, Q. Bagging-based system combination for domain adaptation. *Proceedings of MT Summit XIII, Xiamen, China*, 2011.

[16] Steinbach, M., Karypis, G., and Kumar, V. (2000, August) A comparison of document clustering techniques. *In KDD workshop on text mining* Vol. 400, pp. 525-526.

[17] Stojanovic, N. Ontology-based information retrieval: methods and tools for cooperative query answering (*Doctoral dissertation*, Karlsruhe, Univ., Diss., 2005).

[18] Tiwari, R., and Singh, M. P. Correlation-based attribute selection using genetic algorithm. *International Journal of computer Applications* 4(8), 2010, pp.8875-8887,

[19] Tuominen, J., Kauppinen, T., Viljanen, K., andHyvönen, E. (2009, May). Ontology-based query expansion widget for information retrieval. In Proceedings of the 5th Workshop on Scripting and Development for the Semantic Web (SFSW 2009), 6th European Semantic Web Conference (ESWC 2009) (Vol. 449).

[20] Vafaie, H., and Imam, I. F. (1994, March). Feature selection methods: genetic algorithms vs. greedy-like search. In Proceedings of International Conference on Fuzzy and Intelligent Control Systems.

[21] Wiratunga, N., Koychev, I., and Massie, S Feature selection and generalization for retrieval of textual cases, *In Advances in Case-Based Reasoning Springer Berlin Heidelberg* 2004, pp. 806- 820.