Sophia: A novel approach for Textual Case-based Reasoning

David Patterson[1], Niall Rooney[1], Vladimir Dobrynin[2], Mykola Galushka[1]
[1]Northern Ireland Knowledge Engineering Laboratory
University of Ulster, Jordanstown, BT37OQB, U.K.
{wd.patterson, nf.rooney, mg.galushka}@ulster.ac.uk
[2]St Petersburg State University
198904 Petrodvoretz, St Petersburg, Russia
vdobr@oasis.apmath.spbu.ru

Abstract

In this paper we present a novel methodology for textual case-based reasoning. This technique is unique in that it automatically discovers case and similarity knowledge, is language independent, is scaleable and facilitates semantic similarity between cases to be carried out inherently without the need for domain knowledge. In addition it provides an insight into the thematical content of the case-base as a whole, which enables users to better structure queries. We present an analysis of the competency of the system by assessing the quality of the similarity knowledge discovered and show how it is ideally suited to case-based retrieval (querying by example).

1 Introduction

Textual Case-based reasoning (CBR) is very different in many respects from more conventional CBR applications where case knowledge is usually more easily acquired, structured and adequately (and often simplistically) represented as simple vectors or objects. In textual CBR (TCBR) the knowledge to be embodied within a case is much more intricate in that it contains complex linguistic terms and concepts on various topics which are often encompassed within the same case. Acquiring and representing this knowledge adequately, without loosing its meaning and with a low knowledge engineering overhead to users, remains a challenging prospect. Equally challenging is the prospect of discovering, implementing and maintaining useful similarity knowledge within such systems where it is often vital to identify similar cases that perhaps do not necessarily contain the same words but semantically similar themes or concepts. It may initially seem that the problem of TCBR could be adequately addressed by standard Information retrieval (IR) techniques. However there are limitations with these approaches. Perhaps the best known IR approach, commonly known as the bag of words, transforms documents into a vector space model [Sebastiani, 2002], whereby the case knowledge representation formalism is a vector of weights representing the individual words present in each document and similarity is based on a simplistic comparison of overlap among case vectors. However, as is clear, word order and negation information is lost in the transformation, which may be vital for reasoning in certain domains of interest. Additionally, word sense disambiguation is lost in the process e.g. the word "Java" can have different meanings depending on its context of usage. Most importantly of all, there is a loss in transparency to users, which is a major drawback during case retrieval. NLP techniques can determine word order but they can be brittle and computationally expensive. Some researchers in the CBR community have followed this approach and developed novel but often domain specific approaches to the task of TCBR. Examples of such systems include work by Brüninghaus & Ashley, [2001], Cunningham et al., [2004], Kunze & Hubner [1998] and Lenz [1998]. These systems are very domain-dependent, require considerable knowledge engineering effort and are usually designed for the situation where all documents have a similarly structured content.

Other researchers have instead focused on advancing IR techniques using technologies such as Latent Semantic Indexing [Deerwester et al., 1990] and Probabilistic Latent Semantic Indexing [Hoffman, 1999]. These go beyond the bag-of-words model providing a more advanced view of the document space that may require less knowledge engineering activity within a TCBR context [Zelikovitz and Hirsh, 2002]. Unfortunately these have crucial drawbacks also. For example they use a transformation to project documents into cases and as such they become non transparent to a user. Transparency is an important component of CBR systems to enable the user gain insight into the reasoning and explanation processes of the system. As such they provide little insight into the knowledge contained within a case or how similarity is determined during retrieval. A further limitation is that these approaches are essentially dimensionality reduction techniques and as such important information may

be lost. Additionally word order and negation knowledge is also lost. Wiratunga et al., [2004], combine IR and Machine Learning techniques in their approach. A strength is its ability to facilitate automated semantic similarity determination but unfortunately their approach is completely supervised and as such the class of each document within the collection must be known apriori to enable the discovery of similarity knowledge. This limits the application of the technology. In addition the effectiveness of the technique to multi class domains has not been addressed and remains an open question. Cunningham et al., [2004], have also recently proposed a very different approach to TCBR based on graph theory, that maintains the original document structure and word order. A disadvantage of their approach is that they require an expert to identify domain dependant indexes, which they rely on to assess case similarity using graph distance techniques. Therefore there is a significant knowledge engineering burden in terms of similarity knowledge. In addition it cannot determine semantic similarity between cases or cope with ambiguities that may occur. One final open question with this technique is its efficiency with medium to large case-bases as assessing sub graph similarity is a complex matter.

In this research, we present a new approach for TCBR called SOPHIA CBR, based upon a scaleable contextual document clustering approach [Dobrynin et al., 2004], which facilitates an advanced and rich knowledge discovery framework for case-based retrieval. SOPHIA's case representation formalism is similar to a classical vector space model except, rather than using word frequencies, it is based on the conditional probability distributions of terms within documents. It then intelligently discovers important contexts within the case-base and organizes cases into one of a large number of clusters which have these contexts as attractors. This produces groups of semantically related cases (i.e. cases which are on the same or similar subjects but use different terminology, can be recognized as similar) for a given cluster context. This process of forming clusters, allows both a very efficient and competent case retrieval process. It is this unique feature that provides much of the power behind the technology.

SOPHIA CBR is advantageous in that, it is domain independent, and has low knowledge engineering overheads as it does not require any user intervention to acquire domain knowledge. As such, all knowledge can be discovered automatically (although if background knowledge is already present it can be utilized). Additionally it uses a transparent case knowledge representation, automatically discovers and provides users with additional knowledge about the domain that they can use to refine queries, is language independent, is scaleable and can differentiate between the different contexts of potentially ambiguous terms. The novel technology presented in SOPHIA CBR is useful for both classification tasks and for retrieval, browsing and searching by example. SOPHIA does not have a mechanism to identify word order or negation, features which undoubtedly are important for document collections where each document has a similar internal structure. However in terms of its application to document collections such as presented in this paper (where each document does not necessarily have a similar internal content structure), we show that SOPHIA is capable of providing a competent TCBR system. Firstly we present the SOPHIA technology, then we carry out an initial experiment to demonstrate the quality of the discovered case and similarity knowledge. This is followed by an additional experiment which investigates the potential of the system to case-based retrieval (query by example). Finally we discuss the results and present future work.

2 SOPHIA CBR Methodology

Knowledge is automatically discovered at various stages within the SOPHIA Framework. In step 1 we describe how case knowledge is discovered and represented. In step 2 we show how numerous specific narrow context words are automatically identified. These narrow contexts act as attractors for clustering cases and this step can be regarded as *global* similarity knowledge discovery. In step 3 *cluster level* similarity knowledge is discovered and used to determine which narrow context each case should be assigned to. Step 4, is strictly not part of the clustering algorithm itself but is an additional processing step that provides extra knowledge about the internal case structure of clusters and provide a means for visualizing them. This can be regarded as *localized* similarity knowledge discovery. As will become apparent, not only does this empower the user with extra domain knowledge about the problem area but it improves both the case index and the ability of the system to identify similar cases.

Step 1 Case knowledge Discovery. Here we describe how case knowledge is automatically extracted from a document corpus. In the following definitions, a term refers to a word in the document corpus. Let Ξ denote the set of all documents in the document corpus and Ψ denote the set of all terms present in Ξ. For every document in the corpus a case is automatically extracted and represented by a probability distribution over all terms occurring in that document.

$$p(y \mid x) = \frac{tf(x,y)}{\sum_{t \in \Psi} tf(x,t)}$$

where $tf(x,y)$ is the term frequency of the term y in document x and t is a term from the set of all terms present in the document collection. Although by itself this does not provide a richer case representation than using conventional IR approaches, it does facilitate the process of grouping, indexing and retrieving semantically similar cases, which forms the centerpiece of the power of this technology.

Step 2 Global Similarity Knowledge Discovery. Given a term $z \in \Psi$, we define its context as the probability distribution of a set of words which co-occur with the given term. More specifically the context of the term z is represented in the form of a conditional probability distribution $p(Y \mid z)$, where the random variable Y takes values from Ψ and $p(y \mid z)$ is equal to the probability of randomly selecting the term y in a randomly selected case within which the term z co-occurs. We can approximate this distribution as:

$$p(y \mid z) = \frac{\sum_{x \in \Xi(z)} tf(x,y)}{\sum_{x \in \Xi(z), t \in \Psi} tf(x,t)}$$

where $tf(x,y)$ is the term frequency of the term y in case x and $\Xi(z)$ is the set of all cases from the corpus which contain the term z. It is obvious that in most cases the context of the term z is too general in scope to present useful information about the corpus. So we are interested only in identifying narrow context terms z. The narrowness of the term z is estimated by the entropy of the probability distribution $p(Y \mid z)$:

$$H(Y \mid z) = -\sum_{y} p(y \mid z) \log(p(y \mid z))$$

Let $\Psi(z)$ denote the set of all different terms from cases in $\Xi(z)$. When there is a uniform distribution of terms from $\Psi(z)$ the entropy $H(Y \mid z)$ is equal to $\log \mid \Psi(z) \mid$. According to Heaps Law $\log \mid \Psi(z) \mid = O(\log(\mid \Xi(z) \mid))$ [Baeza-Yates & Ribeiro-Neto, 1999] there is a relationship between the case frequency $df(z) = \mid \Xi(z) \mid$ of the term z and the entropy of its context. To allow for this dependency, we divide the whole set of words into *r* disjoint subsets:

$$\Psi = \bigcup_i \Psi_i$$

$$\Psi_i = \{z : z \in \Psi, df_i \le df(z) < df_{i+1}\}$$

$$i = 1..r$$

Here the threshold df_i satisfies the condition $df_{i+1} = \alpha df_i$ where $\alpha > 1$ is a constant. Choosing narrow word contexts are based on the assumption that in total there are N narrow word contexts. For every $i = 1,..,r$ a set $Z_i \subset \Psi_i$, is selected such that

$$\mid Z_i \mid = \frac{N. \mid \Psi_i \mid}{\sum_{j=1,..r} \mid \Psi_i \mid}$$

and $z_1 \in Z_i, z_2 \in \Psi_i - Z_i \rightarrow H(Y \mid z_1) \le H(Y \mid z_2)$. Then $Z = \bigcup_i Z_i$, where Z is the set of selected narrow contexts. These contexts form the seeds for clustering semantically related cases, where cluster membership (similarity) is measured using the Jensen-Shannon (JS) divergence [Lin, 1991]. In this respect contexts are regarded as global similarity knowledge.

Step 3 Cluster Level Similarity Knowledge Discovery. Narrow contexts $\{p(Y \mid z)\}_{z \in Z}$, discovered in step 2, are considered as cluster attractors. Within this study all cases are grouped into at most one cluster based on the case similarity to the attractor, i.e. this is a hard clustering approach but equally a softer approach could be applied. In this way cases are associated with the contexts they most closely match to form clusters of cases that are on similar (closely related) subjects or themes. The similarity between a case x and the context for the term z is estimated by the JS divergence between the probability distributions p_1 and p_2 representing the case and the context respectively:

$$JS_{\{0.5,0.5\}}[p_1, p_2] = H[\overline{p}] - 0.5H[p_1] - 0.5H[p_2]$$

where $H[p]$ denotes the entropy of the probability distribution p and $\overline{p}$ denote the average probability distribution $= 0.5p_1 + 0.5p_2$. A case x is therefore assigned to a cluster with attractor z if,

$$z = \arg\min_{t \in Z} JS_{\{0.5,0.5\}}[p(Y \mid x), p(Y \mid t)]$$

in other words a case is assigned to the cluster whose attractor it has the highest semantic similarity to.

Once these three stages are completed, we have discovered all case knowledge, all narrow contexts within the case-base and assigned cases to the context they are most semantically similar to. In an equivalent fashion, the similarity between cases within a cluster can be discovered using the JS divergence. As such the lower the JS divergence, the higher the similarity and as will be seen in step 4, it is this similarity knowledge that forms the key to discovering semantically related cases.

Step 4 Localized Similarity Knowledge Discovery. Up to this point all knowledge discovered has been at the global/cluster level. In this step we discover localized similarity knowledge that defines the inner case structure of each cluster. We represent the relationships (similarities) between cases by a graph where each vertex in the graph represents a case. Any two vertices are connected by an undirected edge, whose weight denotes the distance between corresponding cases. This weight is determined as before using the JS divergence. The standard Kruskal's algorithm is used to find the minimum spanning tree (MST) which spans all graph vertices and has the minimum total weight for its edges. The knowledge within the MST can then be presented to a user as a complete description of the internal structure of a cluster relating to a narrow context. Useful knowledge includes,

cases nearer to the top of the tree are most similar to the narrow context, while those further away are less similar. Cases in close proximity within the MST are more similar than cases that are far apart. Cases in one branch of the tree are more similar than those in separate branches. This localized similarity knowledge can either be used to interactively browse the relevant cluster structure looking for useful cases or as a means of accurately classifying a new case.

3 Experiments

In this work we demonstrate the efficacy of the SOPHIA CBR system to textual case retrieval. In the first experiment we investigate the quality of the case and similarity knowledge by demonstrating the high degree of (semantic) similarity between cases within clusters. In the second, we investigate the quality of the retrieval process itself and show how it is ideally suited to case-based retrieval.

3.1 Case base description

In our experiments we use the well known Modified Apte ("ModApte") split of the Reuters-21578 collection (davidlewis.com/resources/testcollections/reuters-21578) containing 9,603 training documents and 3,299 test documents. Not all documents are actually assigned a category. Therefore we only use those that have at least one topic (7,775 from the training set and 3,019 from the test set) in our experiments. All other documents were considered as background information. We use these documents to accumulate information about the document corpus. All training and background documents were used for case knowledge and similarity knowledge discovery (clustering & MST formation). Test documents were used as queries in the retrieval experiments only. It should be noted at this point that the majority of research carried out with this corpus in the past (mostly within the Information Retrieval community) has tended to focus on evaluations using the 10 most popular categories only [Sebastiani, 2002]. As such our task is much more challenging in that we attempt to use all 120 categories to evaluate our technique. The experiments described in this section could equally be applied to other document collections as SOPHIA CBR is domain and language independent.

Initially documents were parsed (transformed to cases) by converting all words from the title and body of a document into lower case, deleting stop-words using a standard stop-words list (SMART, 571 words) and using the Porter algorithm for stemming. We consider a word as any maximal sequence of symbols which start with a symbol in the range a-z, and ends with any symbol between a and z or any symbol between 0 and 9 and in between contains symbols from the set {a-z0-9_-/}. It should be noted that no additional information about the document set is used, apart from the title and body of the documents themselves. In particular, expert assigned knowledge such as category labels are not used in the clustering/retrieval process and as such the approach is totally unsupervised in nature and therefore has as a consequence, no manual knowledge engineering overheads. In total, 38,088 different terms remained after parsing.

3.2 Experiment on similarity knowledge quality

In the first experiment we use the topic categories of cases as a means of independently assessing the quality of our similarity knowledge and hence our clustering process. Note that this topic categorization domain knowledge is *not* utilized as part of the process of case similarity determination (which is based solely on the textual content of cases) but as a means for *assessing* the effectiveness of the discovered similarity knowledge only. This experiment provides us with insight into the quality and meaningfulness of the SOPHIA TCBR clustering process. Our hypothesis is that if the system identifies cases as semantically similar then there should be a high probability that they share many of the same categories. To facilitate this we determine the degree of overlap between adjacent cases in the MST. We therefore evaluate the actual similarity of 2 cases *a* and *b* based on the similarity of the topics assigned to them by experts. Let *T(x)* be the set of topics assigned to case *x* by an expert. To evaluate the similarity between *T(a)* and *T(b)* we use the Jaccard coefficient (JC).

$$JC = \frac{|T(a) \bigcap T(b)|}{|T(a) \bigcup T(b)|}$$

A coefficient of 1 indicates that both cases have identical topics assigned to them and have maximum similarity, while a value of 0 signifies no overlap between topics, indicating minimum similarity. Figure 1 shows the results of this experiment. The most striking observation from this graph is that the vast majority of cases (5796 pairs –75%) have an extremely high JC (0.9-1.0). This provides extremely strong evidence for the high quality of our similarity knowledge and confirms that semantically similar cases, i.e. those linked in the MST, have a large degree of overlap in their categories and therefore must of necessity be genuinely very similar. This also confirms that forming clusters based on narrow contexts combined with a MST based on JS divergence is a powerful approach for determining textual case similarity which also provides real meaning and transparency to users. That is users can easily see the context of all cases within a cluster and also which cases are most similar to others within the cluster. At the opposite end of the graph, where we can observe case pairs with poor category overlap, it can be seen that there are 1472 pairs of cases (19%) with very poor JC and almost no overlap between their categories. An interesting phenomenon can be observed at a JC value of 0.5-0.6. Here we see a slight rise in the number of case pairs. The reason for this is that many cases have only 2 categories and this peak represents the situation whereby they agree on one category and disagree

on the other (e.g. one case could have 2 categories and another only one, which they share).

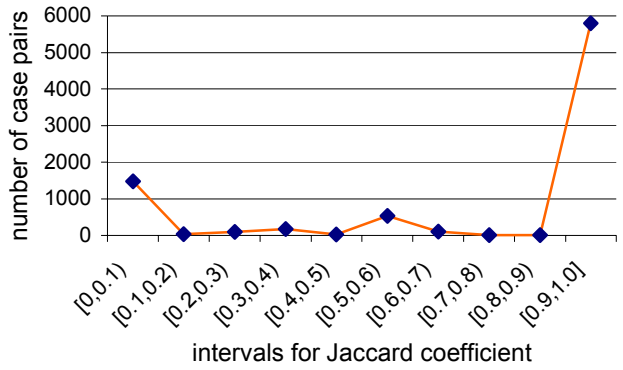

Figure 1 Assessment of similarity for MST edges Using JC

For all other JC values there are practically no case pairs. It is reasonable to conclude that clusters are not entirely homogeneous in terms of their topics. That is, each cluster may contain cases on different topics. Therefore it is to be expected that, within the MST when traversing from one case to another, topic shifts should be encountered. Whenever this occurs adjacent cases in the MST will differ on their topics. These shifts may be gradual, as evidenced by the rise in the graph around a JC of 0.5 (ie adjacent cases overlap on 50% of their topics) or radical, as evidenced by the rise at a JC of 0 (adjacent cases overlap on none of their topics). These results provide compelling evidence for quality of the case and similarity knowledge discovered and utilized by SOPHIA CBR.

3.3 Experiment on Case Retrieval

In this section, we propose a case retrieval approach designed to provide flexible querying plus a high retrieval accuracy and a good explanatory facility. SOPHIA enables queries to be generated using the traditional key word approach or by using query by example, where entire documents (or parts thereof) can be used as the query. SOPHIA then supplies the user with quality knowledge about "all" possibly relevant cases within the case base. It is important that this knowledge should be presented to the user in the context of the whole case collection. In other words, the user should have the facility to estimate:

- which region of the whole case base contains the most relevant cases,
- how large this region is,
- which semantically similar documents are relevant

We propose that the MST is ideally suited to presenting this knowledge to the user. Through a process of system supported browsing, the user can evaluate the relevancy of different parts of the tree, estimate its size and estimate the topics of neighboring cases. This process is demonstrated in the subsequent experiment where we consider the following indexing and browsing scenario. The user has a target case *d* (test case/query example) and the system presents the MST of the most relevant cluster, based on the closest matching context attractor, for browsing. This tree is the MST of a cluster *C(z(d))* whose context *z(d)* is the smallest distance from the target case (as before we use JS divergence to evaluate distances). The nearest neighbor, *NN(z(d),d)*, to case *d* from within cluster *C(z(d))*, is selected from the MST. The user then starts browsing from this case. We will consider three notions of relevance in this experiment. Let *T(x)* be the set of all topics assigned by the expert to case *x*. Then

Predicate $R_{=}(x,y)$ means that $T(x)=T(y)$

Predicate $R_{\subseteq}(x,y)$ means that $T(x)\subseteq T(y)$

Predicate $R_{\cap}(x,y)$ means that $T(x)\bigcap T(y)\neq\varnothing$

In other words if $R_{=}(x,y)$ then cases *x* and *y* are considered to be very relevant as they have exactly the same set of topics assigned by experts. If $R_{\subseteq}(x,y)$ then the relevance of case *x* to case *y* is slightly weaker than in the previous definition but the user can be sure that all topics of case *x* are also contained in case *y*. If $R_{\bigcap}(x,y)$ then we have the weakest notion of relevance in that both cases share at least one common topic. We say that the target case *d* is successfully matched with an existing case if, in the minimum spanning tree *MST(C(z(d)))* of the cluster *C(z(d))* there exists at least one relevant training case *y* within a distance of *k* edge links from the nearest neighbor case *NN(z(d)*. In this evaluation we include the nearest neighbor (NN) case as part of the evaluation and only consider edges which connect to training cases (edges connecting to background cases are ignored). Let predicate *S1(d,k)* indicate that case *d* is successfully matched with an existing case within distance *k,* where relevance is determined by predicate $R_{=}(x,y)$. Similar predicates *S2(d,k)* and *S3(d,k)* are used when relevance is determined by predicates $R_{\subseteq}(x,y)$ and $R_{\cap}(x,y)$ respectively. Let *P1(k),* (*P2(k)*, *P3(k)*) be the probability that for a randomly selected test case *d,* the predicate *S1(d, k)* (*S2(d, k), S3(d, k)*) is *true.* Figure 2 shows how values of *P1(k)*, *P2(k)* and *P3(k)* depend on *k.* These results show that the knowledge discovered and utilized by SOPHIA CBR, when exploited by the process of interactive browsing, provides an ideal medium for locating and retrieving relevant cases. Using the most stringent definition of relevance, where all cases must have exactly he same topics (P1), it can be seen that 75% of the NN cases themselves are relevant (distance 0). It can be seen that this rises to almost 85% when cases within a vicinity of 3 links are considered. Considering larger vicinities, add little more to the relevancy. Defining relevancy as in P2, a similar picture is observed. This time the NN cases are relevant 80% of the time, rising to 87%

when k=3. Again little benefit is added by increasing k further.

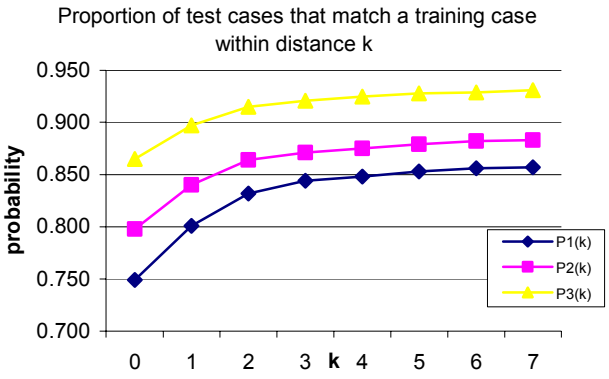

Figure 2 Showing how retrieval relevancy depends on k

Finally the least stringent definition of relevancy provides the best results. 86% of NN cases are relevant, rising to 92% when *k*=3. There is little improvement by increasing *k* further. These results are particularly encouraging, especially when considering that there are 120 possible categories a case can take and many cases have more than 1 category. It should be noted that the least stringent definition of similarity is the truest reflection of similarity within a TCBR system as the goal is to retrieve cases which are semantically similar (ie have some similar topics). This type of retrieval can be regarded as querying by example. A user can cut a piece of text from any other document (sourced from the web for example) and paste it into the SOPHIA system and through the process described retrieve documents that are semantically similar. The authors are unaware of any other scaleable TCBR system (or document clustering algorithm) that can accomplish this.

3 Conclusions

In this paper we present a novel approach for discovering case and similarity knowledge within a TCBR system. We describe how the cases are automatically grouped into semantically related clusters focused around discovered central contexts or themes. We show initially how the case and similarity knowledge discovered is of a very high quality and go on to show how the natural organization of the cases within clusters into a MST, provides a very natural environment to enable case based retrieval (query by example). Important advantages of this technique include the fact it is completely automated, requires no domain knowledge (and therefore no manual knowledge acquisition), it is language independent, can be used for case classification [Dobrynin et al., 2004], facilitates semantic similarity determination and very importantly, unlike all other clustering based approaches for document collections, it is scaleable for very large case bases. This is due to the fact that similarity between cases is only calculated at a local cluster level as opposed to the global level. Future work will include investigating the formation of sub clusters to aid retrieval and case based classification.

References

[Baeza-Yates & Ribeiro-Neto, 1999] Baeza-Yates, R. & Ribeiro-Neto, B. *Modern Information Retrieval*. ACM Press, 1999.

[Brüninghaus & Ashley, 2001] Brüninghaus, S. & Ashley, K., The role of Information Extraction for Textual CBR. In *Proceedings of the 4th International Conference on Case-Based Reasoning*, LNCS 2080, pp. 74-89, Springer, 2001.

[Cunningham et al. 2004] Cunningham, C.M., Weber, R., Proctor, J.M, Fowler, C. & Murphy, M., Investigating Graphs in Textual Case-Based Reasoning. In *Proceedings of the 7th European Conference on Case-based Reasoning*, LNCS, pp. 573-586, Springer, 2004

[Deerwester et al., 1990] Deerwester, S., Dumais, S. T., Furna, G.W., Landauer, T. K. & Harshman, R.. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 1990.

[Dobrynin et al., 2004] Dobrynin, V., Patterson, D. & Rooney, N., Contextual Document Clustering. In *Proceedings of 26th European conference on Information Retrieval Research*, LNCS 2297, pp. 167-180, Springer, 2004.

[Lin, 1991] Lin, J. Divergence Measures Based on the Shannon Entropy. *IEEE Transactions on Information Theory*, 37(1), pp. 145-151, 1991.

[Hofmann, 1999] Hofmann, T. Probabilistic latent semantic analysis. In *Proceedings of the 15 th Conference on Uncertainty in Artificial Intelligence,* pp. 289—296, 1999.

[Kunze & Hubner,1998] Kunze, M. & Hubner, A. CBR on Semi-structured Documents: The Experience Book and the FAllQ Project. In *Proceedings of 6th German Workshop on CBR*, 1998

[Lens, 1998] Lenz, M. Knowledge sources for textual CBR applications. In *Proceedings of the Workshop on Textual Case-Based Reasoning*, pp. 24-29. Menlo Park, CA, AAAI Press. 1998

[Sebastiani, 2002] Sebastiani, F. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1-47, 2002.

[Wiratunga et al., 2004] Wiratunga, N., Koychev, I., & Massie, S. *Feature Selection and Generalisation for Textual Case Retrieval*. In *Proceedings of 7th European Conference on Case-Based Reasoning*, pp. 806-820. LNAI, Springer, 2004.

[Zelikovitz, S. & Hirsh 02] Zelikovitz, S. & Hirsh, H. Integrating Background Knowledge into Nearest-Neighbor Text Classification. In *Proceedings of 6th European Conference on Case-based Reasoning*, pp. 1-5, Springer, 2002.