TCR – Textual Coverage Rate

Kerstin Bach & Alexandre Hanft

University of Hildesheim
Institute of Computer Science
Intelligent Information Systems Lab
D-31141, Hildesheim, Germany
kerstin.bach|alexandre.hanft@uni-hildesheim.de

Abstract

In this paper we will introduce a measure of saturation for unstructured texts of unknown domains. Therefore we will present the Textual Coverage Rate (TCR), a method to determine the IE coverage of unstructured texts using a given vocabulary. We advance efficiency while building vocabulary repositories tailored for given problems and ensure a certain quality of representation. Our approach, which will be evaluated using a large case base, concentrates on the development of the TCR and will motivate its application for textual Case-Based Reasoning.

1 Introduction

In the past many documents were human readable written, but it is still challenging to capture the given information for machines and apply them to new challenges. Before techniques like Case-Based Reasoning (CBR) can be applied for knowledge management, the given data has to be analyzed and prepared. CBR helps to solve problems based on previous experiences and therefore problems and their solutions (cases) have to be analyzed. Furthermore, Textual Case-Based Reasoning (TCBR) offers the opportunity to work with free text documents as well as making previous knowledge and information available. According to [Wilson and Bradshaw, 1999] texts in CBR can be divided in two kinds: fully structured cases and fully textual cases. In this paper we will focus on the fully textual cases which can be separated in text sections of different lengths.

There are many approaches to comprehend textual documents [Asiimwe *et al.*, 2007; Massie *et al.*, 2007] which focus on the formalization and the retrieval of textual documents. In this paper we will present an approach which shows whether the given terms, vocabulary, etc. is comprehensive enough to cover a given case base. We will concentrate on preparing the case bases for TCBR and we will show how this work can be done more efficiently. Also we will show how terms can be extracted and provided to support TCBR.

Before setting up a TCBR-System, for example which is based on Case Retrieval Nets (CRN) [Lenz, 1999], one has to define which kind of data is given in the case base and how it can be accessed. As a result of this preparation a case format is defined, which will contain those information and ensure that the data can be imported and processed adequately. But knowing how to access data does not ensure that the given information can be represented in a proper way. There is a huge amount of information stored in unstructured textual documents which can hardly be processed because it is written in natural language [Quasthoff, 1997] and might contain unknown words. We will show how to minimize the number of unknown words while dealing with data of new, unknown domains and how to cope with texts in natural language. We will give an approach which can be used for small and huge case bases as well.

In the first part of the paper we will describe how a vocabulary repository of terms can be created using heterogeneous data sources. In section 3 we will introduce and explain the elements of the Textual Coverage Rate (TCR) followed of an example which illustrates its calculation, followed by an evaluation of the TCR. The paper will close up with an outline of our future work advancing the TCR.

2 Repositories and Application Data

Before we can introduce the TCR, we have to create a vocabulary repository of terms which can be used to cover texts. Since our application data will be in German we will concentrate on the German language. Also, the procedure can be done for other languages as well. Comparing the English language with the German language there are huge differences of the syntax of inflections. In German, the base form usually differs while building inflections and for that reason we cannot use a stemming algorithm. Instead we decided to build vocabularies containing terms and their inflections which can be used to build CRNs [Lenz *et al.*, 1998].

According to [Lenz, 1999] Information Entities (IE) are terms which are used to build a CRN and each text section is represented of a set of IEs. Therefore the text is divided in text sections regarding to its structure, and the terms contained in a text section which match the given IEs (e.g. of an IE vocabulary) are marked in the CRN. In comparison to full text analysis IEs are easier and faster to determine as long as an vocabulary is available. Hence, the usage of IEs also provides similarity arcs between terms which offers the search for similar terms and the extension of the query (activation of IEs in the CRNs).

2.1 Vocabulary Repository

First we will describe how to create a vocabulary integrating heterogeneous data sources and building up a repository of general terms. According to [Bach, 2007] we have used both, *GermaNet*[1] and a web service of the *Projekt Deutscher Wortschatz*[2] to collect data.

GermaNet is a lexical-semantic net, similar to *WordNet* of the Princeton University, developed at the University of

[1] http://www.sfs.uni-tuebingen.de/lsd/

[2] http://wortschatz.uni-leipzig.de/Webservices/

Tübingen [Lemnitzer and Kunze, 2002]. We have used *GermaNet* to enhance our vocabulary to be able to cover a new domain.

To use the terms and their synonyms in a given vocabulary to build up a case base the *GermaNet* entries have to be integrated in the vocabulary. The terms themselves are used to represent IEs and the semantic relations between terms can be used to assign the similarity arcs. After integrating *GermaNet* we are able to find synonyms on given terms, although *GermaNet* only provides the base forms it covers most of general language terms used in German.

The terms described in *GermaNet* contain no inflections which are important to recognize in natural language texts. Especially in the German language the inflections of a term can differ from its base form. To recognize base terms and inflections the web service provided by the *Projekt Deutscher Wortschatz* can be used, because it is the most comprehensive collection of German words. For each base form the web service returns its inflections which can be stored as terms in the repository and related to its base form.

2.2 Corpus

To evaluate the TCR we will use an application domain of insurance claims consisting of several passages of free text. Each case can be separated in several text sections of different lengths and the vocabulary used to describe the insurance claims is a mixture of general terms and specific term. Therefore we need a vocabulary which covers both and the usage of *GermaNet* gave us a huge amount of general terms so we expect the unknown words to be domain specific terms. The case base contains more than 9.500 cases with 2.2 million words.

A typical case consists of 12 attributes and 8 of them contain unstructured text. In [Bach, 2007] the given 9 attributes are described as retrieval attributes and we consider 6 of them as retrieval attributes free text. So each case we are processing will contain 9 sections and 6 of them will be text sections. How a case is structured and what kind of data it contains can be seen in Table 1, which shows one case and the IEs which match with the dictionary and represent the text section. The table only consists retrieval attributes and the TCR will be only applied for the text sections. Furthermore the column "IE" consists terms which are included in the given text section and can be found in the vocabulary. The IEs found will be used to calculate the TCR as it can be seen in 3.2.

section name	content	IEs
Id	1612	
Ursache	Verdacht auf Flüssigkeitsschaden.VU bittet um Info über Schadensursache und vielleicht über Reparaturmöglichkeiten	Verdacht, Info, Schadensursache, Reparaturmöglichkeiten
Bemerkung	Gerät kann beim Anspruchsteller besichtigt werden.	Gerät
Kurzbeschreibung	HIFI-Stereoanlage	Stereoanlage
Anschaffungswert	200	200
Zeitwert	50	50
Objekt	HIFI- Verstärker Pioneer, Model A- 204R	Verstärker, Model
Zustand	gebraucht, leichte Kratzer am Verstärkergehäuse. An der Front fehlen vier Einstell-Drehknöpfe.	gebraucht, Kratzer, Front, fehlen
Gerätealter	9	9
Schäden	Der Verstärker weist keine spannungsspezifische Funktion auf. Auf der Haupt- und Endstufenplatine sind Flüssigkeitsspuren zu finden. In diesem Bereich sind die Leiterbahnen und Bauteile wie z.B. Widerstände, Kondensatoren und IC-Kontaktbeine korrodiert.	Verstärker, keine, Funktion, Haupt, finden, Bereich, Bauteile, Widerstände

Table 1: An exemplary case (retrieval attributes) of the databased used to evaluate the TCR.

3 Textual Coverage Rate

Most of the domain models of TCBR systems are hand written or adapted from previous applications [Minor, 2006a; Lenz *et al.*, 1998; Hanft and Minor, 2005]. Evaluating whether the existing domain model covers enough terms to represent the given text adequately is challenging. Especially when dealing with a large amount of text an automatic evaluation is needed. Therefore we will introduce a measure of the coverage a given dictionary provides to represent an unknown text. Its name is TCR, Textual Coverage Rate, and it aims at ensuring a higher quality of unstructured text representation.

In [Bach and Hanft, 2007] an approach has been introduced where text sections are analyzed to figure out which words are not described while using a given dictionary representing an unknown text. In the first step all stop words from a given corpus are eliminated because they have no useful information content. As a second step all words which are contained in the available vocabulary are removed and as a result we get a list of words the system does not know is displayed. The knowledge engineer has to model those unknown words to assure a satisfying representation of the text.

Handling large databases the amount of unknown words increases, for this purpose it is important to know which words have to be modeled in first place. In the approach mentioned above lists ordered by frequency have been applied. The impact of modeling words with a high frequency in the source data in the beginning is comprehensive. Like it is shown in Figure 1 the number of unknown words decreased constantly and after looking over the result it has been noticed that words which would help to cover the unknown text were not always been modeled.

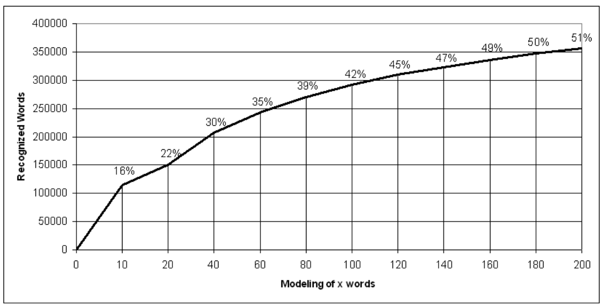

Figure 1: Impact on modeling the 200 most frequent unknown words

3.1 Motivation

Instead of using the frequency of occurrences to prioritize words which have to be modeled first, we will introduce an approach which regards each text section and assigns its coverage with IEs. We aim to indicate words which have to be modeled to ensure each text section is represented satisfactorily. Therefore we focus on both, the expected number of IEs in a given section and the frequency of unknown words.

Preparing an unknown corpus for TCBR requires an analysis if the given dictionary holds suitable terms. We will introduce the Textual Coverage Rate (*TCR*) to describe the potential representation of the source text using the existing dictionary. Therefore we measure the IE coverage of each text section to determine whether it contains a minimum number of terms given in the dictionary or not.

3.2 Calculation

Following [Minor, 2006a] a case c with k text sections can be described as $c = [S^1, S^2, \ldots, S^k]$. Each text section is represented by a set of IEs, called S^i. In addition, we will use T^i which describes the expected number of IEs in each considered text section. In comparison to [Bach and Hanft, 2007] we do not have a global T, because we figured out, that T should depend on the local attribute.

(1) and (2) calculate the number of text sections which contain less IEs than given by T^i.D_{cov} describes the coverage rate of one text section. It is 0 if there are less than T IEs in the tested section S^i. For example a text section is represented by two IEs ($|S^i| = 2$) and three IEs are expected ($T = 3$) this section is less covered and D_{cov} for the considered section will be 0.

$$D_{cov}(S^i, T^i) = \begin{cases} 0 & , \quad |S^i| < T^i \\ 1 & , \quad \text{else.} \end{cases} \tag{1}$$

To calculate the *TCR* the number of appropriate covered text sections has to be summed up and the ratio between this sum and the total number of sections gives the *TCR*:

$$TCR(c, T^i) = \frac{\sum_{i=1}^{k} D_{cov}(S^i, T^i)}{k}. \tag{2}$$

The *TCR* shown above describes the percentage of text sections represented by at least T IEs. If every text section is adequately covered (for each text section $|S^i| \geq T$ is true) the *TCR* will be 1. Otherwise the knowledge engineer should model more terms to increase the coverage of the given dictionary. To figure out which words should be added to the dictionary the approach described in [Bach and Hanft, 2007] can be used.

Furthermore, if the *TCR* is 1, the percentage of text sections which contain more than T IEs should be calculated. For that reason the ratio of excess coverage can be examined as shown in (3) and (4). In opposite to (1) and (2) only text sections represented by more than T IEs are factored.

$$D_{excess}(S^i, T) = \begin{cases} 1 & , \quad |S^i| > T \\ 0 & , \quad \text{else.} \end{cases} \tag{3}$$

$$C_{excess}(c, T) = \frac{\sum_{i=1}^{k} D_{excess}(S^i, T)}{k}. \tag{4}$$

A high excess coverage ratio C_{excess} (more than 0.8) points out that more than the expected T IEs represent a text section and the knowledge engineer can consider increasing T. After increasing T the *TCR* has to be updated and the recalculated C_{excess} helps to decide whether T is chosen correct or still too low. If necessary this step has to be repeated until a C_{excess} of 0.5 or less occurs.

The *TCR* can be used to explain to the knowledge engineer how many words have to be modeled to achieve a certain quality (given by T) covering the corpus. In addition the C_{excess} can increase the quality of coverage, because it shows how many words have to be modeled to increase T.

3.3 Calculation Example

According to the given data in Table 1, Table 2 shows in detail how the TCR is calculated. S^i counts the given sections which will be referenced during the calculation. The column $|S^i|$ contains the number of IEs found in the text section and according to section 3.2 D_{cov} is calculated. T is chosen according to the average number of IEs found in each section. k contains the total number of text sections and the TCR in the given example is 0.5. To consider increasing T, C_{excess} is calculated based on the given data. The result 0.3 shows, that most of the considered text sections to calculate C_{excess} are represented of T IEs.

	S^i	$\lvert S^i \rvert$	T^i	D_{cov}	D_{excess}
Id					
Ursache	1	4	1	1	1
Bemerkung	2	1	1	1	0
Kurzbeschreibung	4	1	2	0	0
Anschaffungswert					
Zeitwert					
Objekt	8	2	7	0	0
Zustand	9	4	1	1	1
Gerätealter					
Schäden	11	8	10	0	0
Σ				3	
	$k = 6$		$TCR = 0.5$		$C_{excess} = 0.3$

Table 2: Calculation of the TCR based on the exemplary case of Table 1.

As a result the solutions show that the knowledge engineer should first model terms contained in sections which were rated 0 while calculating D_{cov}. Hence, T was chosen adequate for each attribute to represent the given text sections. As described before we chose T according to the average number of IEs in each section.

4 Evaluation

After introducing the TCR now we will present the evaluation using the previously described corpus (see section 2.2). For the evaluation we use a corpus of 9640 cases and each case looks like the example given in Table 1. In a first step we calculated the average number of IEs occurring in each section. The results are given in in Table 3.

	S^i	min	max	avg	T^i
Ursache	1	0	20	1.16	1
Bemerkung	2	0	20	0.71	1
Kurzbeschreibung	4	0	12	2.20	2
Objekt	8	0	53	7.18	7
Zustand	9	0	32	1.50	1
Schäden	11	0	71	9.98	10

Table 3: Summary of IE distribution

Table 3 shows that depending on the text section the number of IEs found can vary. To assign an appropriate T

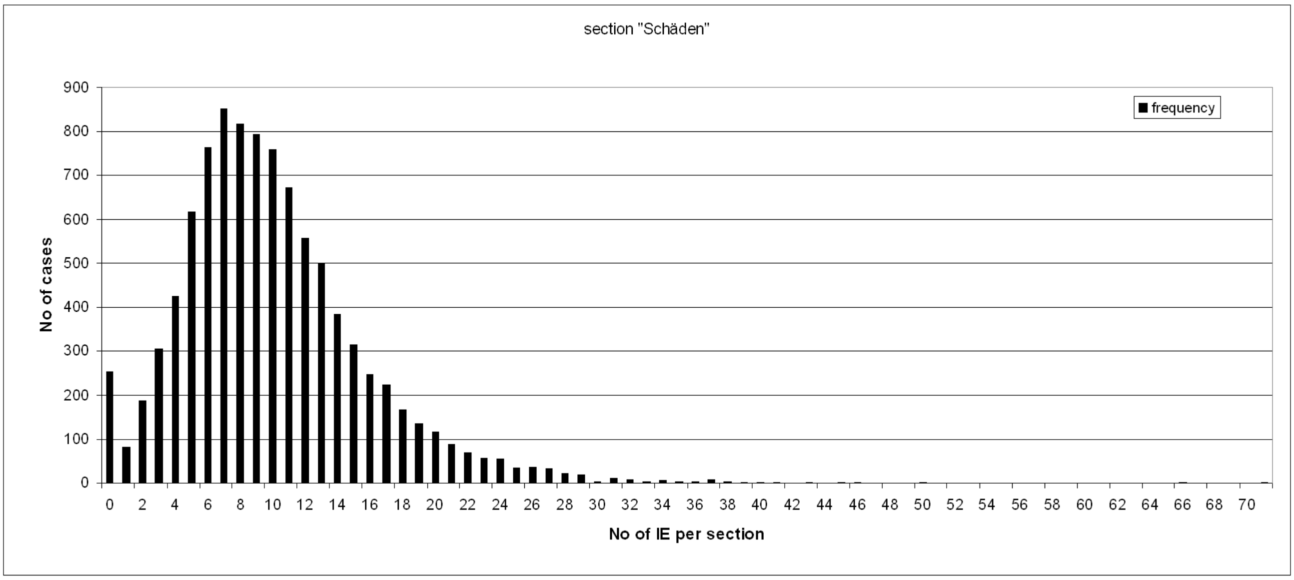

Figure 2: Histogram of found IEs per section

we first choose the average number of IEs per section which can later be refined. As an advancement in comparison to [Bach and Hanft, 2007] we do not have a global T, because during evaluation we figured out, that the variance can be large. For example in text section 4, ”Kurzbeschreibung”, and text section 11, ”Schäden”, we found a minimum of 0 IEs. ”Kurzbeschreibung” has a maximum of 12 IEs unlike ”Schäden” contains a maximum 71 IEs. This difference would make it hard to assign a global TCR while we are aiming for an adequate representation of each text section.

In a first step we recommend to choose the average number of IEs to figure out how the text can be represented. Depending on the given data the T might have to be increased. For the actual case base we get the following results (which can be seen in Figure 2) for the attribute ”Schäden” after choosing the average number of IEs per text section.

Figure 2 shows the frequency of the IEs found in each text section of ”Schäden”. Although the average number of IEs is 10, more than 50% of the text sections contain less IEs which can be seen as positive skew in Figure 2. The skew for the distribution of the section ”Schäden” is 1.30. We computed the frequency distribution of the other five sections as well and all of them show a similar positive skew.

Obviously the much higher the value is chosen for T, the number of IEs which have to be modeled increase, because the TCR of more cases is insufficient. Now the knowledge engineer has to decide whether the given coverage is exhausted or if a modeling is necessary. In our example, for text section ”Schäden”, it is required to remodel terms, because the text section contains a lot of information and the retrieval is going to need this data to work properly.

Changing T influences the saturation of the text sections of IEs, which assures a higher probability of retrieving correct documents. But aiming at a high T means that many terms have to be modeled which might be time-consuming. Using the TCR can tell the knowledge engineer how many terms have to be modeled and he will be able to decide if it is worth it. After evaluating our data using different numbers of cases, we suggest to take the median as value for T.

TCR	No of cases
0	8
0.167	272
0.333	1676
0.5	3339
0.667	2938
0.833	1258
1	149

Table 4: Number of cases with the same TCR

For our test case base, we calculate the TCR for each case and get an average of 0.563. Having a closer look at the distribution and summing up the cases with the same TCR, which can be seen in 4, one third of the cases have a TCR of 0.5. About 45% have a TCR of 0.6 or higher, but although 8 cases have no sufficiently filled sections as well as 1956 cases have more than an half insufficient text sections.

In our example more modeling is required because the considered text section contains a lot of information and the retrieval is going to need this data to work properly.

5 Conclusion & Outlook

This paper devotes furthering the performance of TCBR applications concentrating on the preparation of data and case base. We described the usage of *GermaNET* and *Projekt Deutscher Wortschatz* to create a vocabulary of terms which will be used represent unknown texts. Creating vocabularies from scratch is challenging, so we based our vocabulary on the ExperienceBook II[3] vocabulary, which was developed at the Humboldt University of Berlin [Hanft and Minor, 2005; Minor, 2006a; 2006b]. Nevertheless, the presented approach using *GermaNET* and *Projekt Deutscher Wortschatz* can be used to build a new vocabulary as well. Furthermore, the English complement of *GermaNET*, *WordNET*[4] is also available and can be used just like the German version.

[3]https://roy.informatik.hu-berlin.de/ExpBookII/

[4]http://wordnet.princeton.edu/

The Textual Coverage Rate (TCR) presented in this paper measures the coverage of IE of unstructured text section using a given vocabulary and facilitate a deep insight in the considered text corpus. This empowers the knowledge engineer to decide which parts of the corpus should be modeled first and how much should be done to achieve a certain quality of modeling which means coverage of unstructured text through IEs. The TCR was evaluated using a corpus with over 9500 cases.

Similar work is done at Robert Gordon University. [Massie *et al.*, 2007] describes the extraction features from text in anomaly reports to map them to structured cases. As in our approach, [Wiratunga *et al.*, 2005] also motivate the pre-processing of data to extract features, but this work uses rules to extract features, like the Propositional Semantic Indexing (PSI). In comparison to our approach, PSI depends on the domain and relies on word-class cooccurrences. Parts of the calculation of TCR are like the vector space model [Salton *et al.*, 1975], especially the term frequency, but TCR avoids the problem of small similarity values by long documents and, as usually in CBR, the requirement of an exact matching of the key words in query and case is not necessary.

In future work we will use TCR to facilitate automatic maintenance of knowledge in a Case Factory, which has been described in [Althoff *et al.*, 2006]. Furthermore we will concentrate on developing the TCR aiming at its application in systems based on CoMES [Althoff *et al.*, 2007], because those applications will for example process contributions on community platforms.

Currently, our vocabulary repository contains terms and their synonyms, but we are aiming at enhancing the vocabulary with similarities between terms, so CRNs can demonstrate their strength dealing with fully textual cases.

Another challenge for the future will be coping with homonyms which can possibly be done using a semantic classification of terms to figure out the meaning of a certain term. Therefore the classes of the terms in the text section have to be considered and according to those we might be able to assign the homonyms' meaning.

References

[Althoff *et al.*, 2006] Klaus-Dieter Althoff, Alexandre Hanft, and Martin Schaaf. Case Factory – Maintaining Experience to Learn. In Mehmet H. Göker, Thomas Roth-Berghofer, and H. Altay Güvenir, editors, *Proc. 8th European Conference on Case-Based Reasoning (ECCBR'06), Ölüdeniz/Fethiye, Turkey*, volume 4106 of *Lecture Notes in Computer Science*, pages 429–442, Berlin, 2006. Springer Verlag.

[Althoff *et al.*, 2007] Klaus-Dieter Althoff, Kerstin Bach, Jan-Oliver Deutsch, Alexandre Hanft, Jens Mänz, Thomas Müller, Régis Newo, Meike Reichle, Martin Schaaf, and Karl-Heinz Weis. Collaborative Multi-Expert-Systems – Realizing Knowlegde-Product-Lines with Case Factories and Distributed Learning Systems. In Joachim Baumeister and Dietmar Seipel, editors, *Accepted for the Workshop Proceedings on the 3rd Workshop on Knowledge Engineering and Software Engineering (KESE 2007)*, Osnabrück, Germany, 2007.

[Asiimwe *et al.*, 2007] Stella Asiimwe, Susan Craw, Bruce Taylor, and Nirmalie Wiratunga. Case Authoring: from Textual Reports to Knowledge-rich Cases. In Michel Richter an Rosina Weber, editor, *ICCBR07*, volume 4626 of *Lecture Notes in Artificial Intelligence*, pages 179–193, Berlin, 2007. Springer.

[Bach and Hanft, 2007] Kerstin Bach and Alexandre Hanft. Domain Modeling in TCBR Systems: How to Understand a New Application Domain. In David C. Wilson and Deepak Khemani, editors, *Proceedings of the 7th International Conference on Case-Based Reasoning (ICCBR) 2007, Workshop on Knowledge Discovery and Similarity*, pages 95–103, Belfast, Northern Ireland, 2007.

[Bach, 2007] Kerstin Bach. Domänenmodellierung im Textuellen Fallbasierten Schließen. Master's thesis, Institute of Computer Science, University of Hildesheim, 2007.

[Hanft and Minor, 2005] Alexandre Hanft and Mirjam Minor. A Low-Effort, Collaborative Maintenance Model for Textual CBR. In Stefanie Brüninghaus, editor, *ICCBR 2005 Workshop Proceedings*, pages 138–149, 2005.

[Lemnitzer and Kunze, 2002] Lothar Lemnitzer and Claudia Kunze. GermaNet - representation, visualization, application. In Rodriguez, M.G., and C.P.S. Araujo, editors, *Proceedings Conference on Language Resources and Evaluation (LREC) 2002, main conference, Vol V.*, pages 1485–1491, 2002.

[Lenz *et al.*, 1998] Mario Lenz, André Hübner, and Mirjam Kunze. Textual CBR. In Mario Lenz, Brigitte Bartsch-Spörl, Hans-Dieter Burkhard, and Stefan Wess, editors, *Case-Based Reasoning Technology – From Foundations to Applications*, Lecture Notes in Artificial Intelligence, LNAI 1400, pages 115–137. Springer-Verlag, Berlin, 1998.

[Lenz, 1999] Mario Lenz. *Case Retrieval Nets as a Model for Building Flexible Information Systems*. Dissertation, Humboldt University of Berlin, Berlin, 1999.

[Massie *et al.*, 2007] Stewart Massie, Nirmalie Wiratunga, Alessandro Donati, and Emmanuel Vicari. From Anomaly Reports to Cases. In Michel Richter and Rosina Weber, editors, *ICCBR07*, Lecture Notes in Artificial Intelligence, pages 359–373. Springer, 2007.

[Minor, 2006a] Mirjam Minor. *Erfahrungsmanagement mit fallbasierten Assistenzsystemen*. Dissertation, Humboldt University of Berlin, May 2006.

[Minor, 2006b] Mirjam Minor. Experience Management with Case-Based Assistant Systems. In Thomas Roth-Berghofer, Mehmet H. Göker, and H. Altay Güvenir, editors, *ECCBR 2006*, volume 4106 of *Lecture Notes in Artificial Intelligence*, pages 182–195, Berlin, 2006. Springer.

[Quasthoff, 1997] Uwe Quasthoff. Projekt Der Deutsche Wortschatz. In Gerhard Heyer and Christian Wolff, editors, *GLDV-Jahrestagung*, pages 93–99. Deutscher Universitäts-Verlag, 1997.

[Salton *et al.*, 1975] Gerard Salton, A. Wong, and C. S. Yang. A Vector Space Model for Automatic Indexing. *Communication of the ACM*, 18(11):613–620, November 1975.

[Wilson and Bradshaw, 1999] David C. Wilson and Shannon Bradshaw. CBR Textuality. In Stefanie Brüninghaus, editor, *Proceedings of the Fourth UK Case-Based Reasoning Workshop*, pages 67–80, University of Salford, 1999.

[Wiratunga *et al.*, 2005] Nirmalie Wiratunga, Robert Lothian, Sutanu Chakraborti, and Ivan Koychev. A Propositional Approach to Textual Case Indexing. In Alipio Jorge, Luis Torgo, Pavel Brazdil, Rui Camacho, and Joao Gama, editors, *PKDD*, volume 3721 of *Lecture Notes in Computer Science*, pages 380–391. Springer, 2005.