Proceedings of the Twenty-Second International FLAIRS Conference (2009)

Reasoning about Changes of Corpus of Documents: Reasoning on Association Rules

Laurent Perrussel
IRIT - Université de Toulouse
2, rue du doyen Gabriel Marty
F-31042 Toulouse, France
laurent.perrussel@irit.fr

Abstract

Evaluating changes in documentation of technical products is a key issue in knowledge management. A product may be declined in different versions and one way to evaluate changes is to compare the sets of documents which describe each version. The aim of this paper is to propose a framework for exhibiting changes between sets of documents. This framework is based on the representation of the sets of documents in terms of association rules and on the definition of first order predicates for reasoning with these association rules. The aim of the reasoning stage is to exhibit the differences between the sets of documents. These predicates show what rules are specific to a corpus or how differs the usage of concepts appearing in the associations rules. The framework is experimented with the comparison of two corpuses of documents which describe documentation about two different versions of a spatial component.

Introduction

A fundamental challenge in the effective handling of large collection of documents, is the ability to identify changes in an efficient way. Products may change over years (software, car, satellite...) and one way to assess the evolution of a product is to compare documentation associated to each version of the product. The aim of this paper is to show a framework for the evaluation of changes based on the comparison of two sets of documents. This framework requires that each set is represented as a set of association rules in order to compare documentation.

The rule extraction stage may produce a very large set of association rules and thus the comparison process should be able to select the association rules that stress the differences. There already exists numerous techniques for evaluating the interest of the discovered association rules: definitions of metrics (Tan, Kumar, & Srivastava 2002), application of templates (Klemettinen *et al.* 1994) or filtering with the help of a knowledge base (Resnik 1995). However all these techniques propose methods for defining what are the "best" or the "most" relevant rules for representing a set of documents. Nevertheless, since they do not consider several corpuses they fail to bring an interesting framework in the context of comparison of sets of documents: a "rule" is interesting because it stresses differences between the sets and thus between knowledge represented by these sets.

For this comparison stage we first define a closeness value between documents; this value characterizes the similarity between documents. Second, we introduce several criteria of relevance for exhibiting relevant association rules. A first criterion of *relevance* of a rule is its specificity aspect: even if two documents which belong to two different sets of documents are similar, they do not support the same association rules. Consequently, rules supported by the documents belonging to only one set of documents stress documentation change. At the opposite, we propose a second criterion of relevance which exhibits the "foundational" shared knowledge in different documentation sets: some rules have been elicited with the help of documents that deeply differ. For all the proposed criteria, the underlying idea is to stress the differences by confronting association rules and their original sources (the sets of documents). For this, we propose to represent information about documents and mined association rules with the help of a knowledge base. This knowledge base expressed in terms of first order predicates enables to perform simple reasoning steps which show documentation change.

In order to evaluate the relevance of the proposed framework, we detail its application in the context of documentation management in the spatial domain. That is, two versions of a component of a spatial satellite are studied through its documentation.

The paper is structured as follows: at first, we detail the process that leads to the representation of documents in terms of association rules. Next, we show how we evaluate the relevance of the rules by introducing first order predicates which aim at stressing differences and closenesses between rules and documents. Next, we show how we analyze the set of association rules with the help of the previously introduced predicates; in this section we also show how to take into account and compare multiple sets of documents. Finally, we evaluate the relevance of the proposed framework with the help of an example. We conclude the paper by outlining some future works.

Building the set of rules

In this section we describe the process that leads to the production of the set of association rules based on a set of raw documents (Feldman & Sanger 2006). As initial input, we consider all textual documents that belong to a collection characterizing a similar domain such as technical documentation. Our purpose is to compare documents or collections of documents and this goal entails that all documents belonging to a collection have to be distinguishable in terms of contents. That is, multiple copies or draft versions of the same document have been removed from the initial collection. Let W be the cleaned collection of documents. Next, we extract the concepts characterizing W. At this stage, two options are available: extracting the concepts by only using documents of W or using an auxiliary knowledge base. Using an auxiliary knowledge base such as a taxonomy enables to improve the quality, i.e. the relevance, of the final set of association rules (Richardson, Smeaton, & Murphy 1994; Resnik 1995). However, it supposes at first that a relevant taxonomy pre-exists, and second this taxonomy has to be independent of the collections that will be compare. That is, if we want to compare two collections of documents which represent two versions of a technical product, the taxonomy has to be independent of these two different versions. In an industrial context, this issue seems quite hard to handle and it has led us to avoid the usage of an external knowledge base. Instead, we propose to extract the concepts from the documents themselves. The proposed technique consists of the usage of classical information retrieval techniques: concepts are extracted from documents and next a weight is associated to each extracted concept. In the following, concepts are defined with the help of the words that appear in the documents with a strong restriction: we limit ourselves to one-word concepts. We enforce this restriction since our main goal is to focus on the evaluation of the association rules and we want to extract concepts in a quick and simple way; however considering multiple words concepts may be done latter. Each document is thus characterized by a list of keywords with a weight calculated as follows (Salton & Buckley 1988):

$$w_{i,j} = \begin{cases} 0.5 \times \left(1 + \frac{p_{i,j}}{\max\limits_{l}(p_{i,l})}\right) \log\left(\frac{|W|}{n_j}\right) & \text{if } p_{i,j} \neq 0 \\ 0 \text{ otherwise} \end{cases}$$

such that $w_{i,j}$ denotes the weight of word i in document j, $p_{i,j}$ is the relative document frequency of word i in document j and n_j is the total number of documents of W where word i appears. In other terms, the words which have the highest weight w.r.t. a document, are the words that appears frequently in that document without appearing too much in the other documents, i.e. these words are the keywords of the documents. This technique has the advantage to eliminate words with a too weak semantic without building a specific thesaurus for them; for instance the word 'satellite' usually appears in almost all documents of a spatial satellite project. It means that we set a threshold γ in order to select the words that have to be considered as the keywords of a document (i.e. the relevant concepts).

In the following, we rewrite the classical functions which enable us to product the association rules (Feldman & Sanger 2006). Function coverage represents the covering set w.r.t. a set of keywords S, threshold γ and collection W:

$$\mathsf{coverage}(S, \gamma, W) = \{j \mid j \in W \text{ and } \text{forall } i \in S, w_{i,j} \geqslant \gamma\}$$

The next stage consists of setting data for the reasoning stage. At first we generate a set of association rules $R : K \Rightarrow k$ w.r.t. a set of keywords entailed by γ:

$$R : \text{keyword}_1, \text{keyword}_2 \cdots \text{keyword}_i \Rightarrow \text{keyword}_n$$

Let $\mathsf{support}(R, W, \gamma)$ be a function which describes the *support* of rule R w.r.t. collection W and threshold γ:

$$\mathsf{support}(R, W, \gamma) = |\mathsf{coverage}(K \cup \{k\}, \gamma, W)|$$

Let $\mathsf{confidence}(R, W, \gamma)$ be the confidence rate of R w.r.t. collection W and threshold γ:

$$\mathsf{confidence}(R, W, \gamma) = \frac{|\mathsf{coverage}(K \cup \{k\}, \gamma, W)|}{|\mathsf{coverage}(K, \gamma, W)|}$$

Finally, w.r.t. some support and confidence thresholds, respectively denoted σ and χ, we obtain the set of association rules:

$$\mathsf{rules}(W, \gamma, \sigma, \chi) = \{R | \mathsf{support}(R, W, \gamma) \geqslant \sigma \text{ and } \mathsf{confidence}(R, W, \gamma) \geqslant \chi\}$$

In order to evaluate the relevance of set of association rules, we first need to calculate the set of keywords that characterize each document j w.r.t. threshold γ:

$$\mathsf{keywords}(j, \gamma) = \{i \mid w_{i,j} \geqslant \gamma\}$$

Second, we evaluate a similarity degree between all documents based on a scalar product and a norm calculus (Salton 1989):

$$\mathsf{similarity}(j, j', \gamma) = \frac{\mathsf{keywords}(j, \gamma) \cdot \mathsf{keywords}(j', \gamma)}{\|\mathsf{keywords}(j, \gamma)\| \times \|\mathsf{keywords}(j', \gamma)\|}$$

In the following, we only use this set of data. However, we use it in a specific way: we propose to reason about the rules and the similarity values. Several works have been done to reason with the help of association rules (such as causal reasoning (Silverstein *et al.* 2000)); in this work, our aim is to define a knowledge base that reflects computed data and next to investigate the knowledge base and thus the corpus by defining predicates. Numerical data are rephrased using knowledge base $K(\sigma, \chi, \gamma)$, i.e. K is relative to thresholds σ, χ and γ:

- `similarity`(j, j', s): the similarity between documents j and j' w.r.t. threshold γ is equal to s s.t. $s = \mathsf{similarity}(j, j', \gamma)$.
- `weightWordDoc`(i, j, w): the weight of word i in document j is equal to w s.t. $w = w_{i,j}$
- `keywords`(i, j): word i belongs to the set given by $\mathsf{keywords}(j, \gamma)$.

- $\texttt{inCorpus}(j, W)$: document j belongs to set W.
- $\texttt{ruleCsq}(r, c)$: keyword c appears on the right part of rule r s.t. $r : A \Rightarrow c \in \mathsf{rules}(W, \gamma, \sigma, \chi)$.
- $\texttt{ruleAnt}(r, a)$: keyword a appears on the left part of the rule r s.t. $r : A \Rightarrow c \in \mathsf{rules}(W, \gamma, \sigma, \chi)$ and $a \in A$.

In the next sections, we show at first how the knowledge base can be used in order to select a subset of relevant rules and, second we exhibit the reasoning steps that lead to exhibit change.

Evaluating the relevance of the rules

The main drawback with the association rules computation is to get a large number of rules where most of the rules are poor from a semantic point of view (i.e. they report obvious relations). Numerous works have been done in order to avoid this problem. A first technique consists of the aggregation of rules (Toivonen *et al.* 1995): by considering an auxiliary taxonomy, rules can be gathered with the help of more general concepts. A second technique consists of evaluating rules; this technique also consists in the usage of an extra knowledge base: in (Feldman & Hirsh 1996), R. Feldman and H. Hirsh suggest to consider extra predicates that must be satisfied by the discovered rules. Both of these techniques entail the availability of an extra knowledge base; as previously mentioned, we do not want to consider extra knowledge, we want to base all the results on the set of documents. Hence, we propose to avoid the problem of semantically weak rules by using knowledge given by inter-documents similarity data. We evaluate the relevance w.r.t. the following criteria:

- what are the rules which have been discovered with the help of documents that are weakly similar?
- what rules are specific? That is, if we suppose two similar documents, can we exhibit an association rule that is only related to one document.

The first question enables to discover knowledge that brings a rich semantic. The documents involved in the definition of the rules are different in terms of content but they relate concepts in similar way. We advocate that this characteristic emphasizes the semantic of the association rules. The second criterion enables to focus on the rules that are "really" specific to some documents. Again, we advocate that this specificity aspect emphasizes the semantic content of the association rules and is helpful for describing documentation change.

These rules can be described in more formal terms as follows. Suppose knowledge base $K(\sigma, \chi, \gamma)$. At first, we want to focus on similar documents, let κ be a threshold with a value that should be closed to 0 in order to evaluate if two documents are deeply different; these documents are characterized by the predicate `diffDocs`.

$$\texttt{similarity}(j, j', s) \wedge (s \leqslant \kappa) \rightarrow \texttt{diffDocs}(j, j', \kappa)$$

Next, we relate documents and rules: what documents have been involved in the production of a rule. In order to establish the relation, we suppose that the consequent and one of the keywords of the antecedent of a rule have to appear in the document. That is formally described using new predicate `rulesDocs`:

$$\begin{aligned}(\texttt{weightWordDoc}(i, j, w) \wedge \texttt{ruleCsq}(r, i) \wedge \\ \texttt{ruleAnt}(r, i') \wedge \texttt{weightWordDoc}(i', j, w')) \rightarrow \\ \texttt{rulesDocs}(r, j)\end{aligned}$$

Using these two inference rules, we are now able to characterize what are the rules shared by two documents which are said to be different. In the following inference rule, we still refer to threshold κ which sets the meaning of our notion of different documents. Predicate `commonRules` is defined as follows:

$$\begin{aligned}(\texttt{rulesDocs}(r, j) \wedge \texttt{rulesDocs}(r, j') \wedge \\ \texttt{diffDocs}(j, j', \kappa)) \rightarrow \texttt{commonRules}(r, j, j', \kappa)\end{aligned}$$

In a similar way, we focus on the differences between two documents that are similar. For this, we first define a rule that characterizes similar documents, the similarity is defined w.r.t. a threshold κ that should be closed to 1 in order to enforce the closeness criterion:

$$\texttt{similarity}(j, j', s) \wedge (s \geqslant \kappa) \rightarrow \texttt{closedDocs}(j, j', \kappa)$$

Next, we define an inference rule which states that if a rule is not shared by two similar documents, then it means that this rule is really specific to one document. In the following, predicate $\texttt{diffRules}(j, j', r)$ stands for rule r is specific to document j compared to document j'.

$$\begin{aligned}\texttt{closedDocs}(j, j', \kappa) \wedge \texttt{rulesDocs}(r, j) \wedge \\ \neg\texttt{rulesDocs}(r, j') \rightarrow \texttt{diffRules}(j, j', r)\end{aligned}$$

It means that we are now in a situation where we can quickly focus on relevant rules. Nevertheless, setting threshold κ is a key issue: in order to define its value and thus setting the meaning of different and similar documents, we suggest the following heuristic. First, we compute similarity values and next, with the help of the experts we isolate a subset of documents that should be similar and use the associated average similarity value as threshold κ. We proceed in a similar way to set the value characterizing the notion of different documents. The next issue is to consider multiple sets of documents and introduce this dimension in the reasoning process.

Evaluating the differences between two sets of Documents

Our aim is to focus on rules that are relevant and that stress the changes between sets of documents. In the following, for the sake of conciseness we only consider two sets of documents (corpuses) that are related to the same topic. At first, we have to suppose that no document belongs to two distinct corpuses otherwise the reasoning stage will be biased. Second, we want to focus on characteristics relating two sets of documents. It means that the set of association rules has to be built by considering all documents at the same time.

We start by focusing on the different documents that belong to distinct corpuses. These documents are described with the help of predicate `diffDocsNoOrigin`:

$$(\texttt{diffDocs}(j,j',\kappa)\wedge\texttt{inCorpus}(j,W)\wedge \texttt{inCorpus}(j',W'))\rightarrow\texttt{diffDocsNoOrigin}(j,j',\kappa)$$

The question related to the documents satisfying predicate `diffDocsNoOrigin` is the following: which association rules are related to those documents. These rules characterize the "stable" part or common knowledge shared by the two sets of documents. It reflects the fact that the way to describe knowledge may have change (different documents) however some underlying knowledge has not changed (shared association rules). That is:

$$(\texttt{diffDocsNoOrigin}(j,j',\kappa)\wedge \texttt{commonRules}(r,j,j',\kappa))\rightarrow\texttt{commonKnowledge}(r)$$

At the opposite, we propose to isolate parts of knowledge specific to each corpus. In other words, if we consider that the two corpuses both represent documentation for a product, the aim is to exhibit the rules that have disappeared or appeared during the evolution of the product, or more precisely its related documentation. At first, we extract from the knowledge base similar documents issued from different corpuses.

$$(\texttt{closedDocs}(j,j',\kappa)\wedge\texttt{inCorpus}(j,W)\wedge \texttt{inCorpus}(j',W'))\rightarrow\texttt{closeDocsNoOrigin}(j,j',\kappa)$$

And second, we extract knowledge that has changed:

$$(\texttt{closeDocsNoOrigin}(j,j',\kappa)\wedge \texttt{diffRules}(j,j',r))\rightarrow\texttt{changedRules}(r)$$

Now, we can exhibit what knowledge has disappeared or appeared by setting which set is the first version of documentation and which one represents the second version. Notice that the previous inference rules only consider association rule ids rather than their content. It means that a drawback of the method is that we cannot evaluate the evolution of the rules; that is, for instance, if the support of an initial rule have been expanded or contracted. We are aware of this limit, however this issue is not so easy to tackle: the definition of the evolution of a rule is still an open issue (change of the support? change of the conclusion?...) and we will not consider it in this paper.

The previous predicates aim at confronting similar and opposite information. However, it may be the case that some association rules may only be related to only one set of documents. This configuration is only relevant if we consider that the rules have been obtained w.r.t. all sets of documents. It means that these rules are specific to a set W since they have been generated by consider all documents at the same time. By specific, we mean that all its components (antecedent and consequent) are keywords which characterize only the documents of a set W. Let us now restate this with the help of the following inference rules; at first we characterize to which set is related the consequence of a rule:

$$(\texttt{inCorpus}(j,W)\wedge\texttt{ruleCsq}(r,c)\wedge \texttt{keywords}(c,j))\rightarrow\texttt{specCsq}(r,W)$$

Second, we characterize a keyword of the left part of a rule:

$$(\texttt{inCorpus}(j,W)\wedge\texttt{ruleAnt}(r,a)\wedge \texttt{keywords}(a,j))\rightarrow\texttt{specAnt}(r,W)$$

Next, we set the link between a rule r and a set W:

$$\texttt{ruleCsq}(r,c)\wedge\texttt{specCsq}(r,W)\wedge \forall a(\texttt{ruleAnt}(r,a)\rightarrow\texttt{specAnt}(a,W))\rightarrow \texttt{specRule}(r,W)$$

This inference rule states that a rule is specific to W by considering the following condition: if the conclusion is related to W then all members of the left part of the rule should also be related to W. It means that we have to take care of the case where a conclusion may appear in two association rules but each left part uses keywords that are specific to one set. The following inference takes care of this case:

$$(\texttt{specRule}(r,W)\rightarrow\forall W'(\texttt{specRule}(r,W')\rightarrow W=W')\rightarrow\texttt{onlyOneSet}(r,W)$$

Predicate `onlyOneSet` only exhibits rules that are specific to one set of documents. In other words, these association rules represent what are the key issues handled in one corpus but which are not important at all in the second one. Suppose that one set represents the documentation for a first version of a product and the second set the second version, the previous inference rule helps to show what information is no longer important in the second version; or at the opposite what is really new.

Illustration

We illustrate the proposed framework by considering two versions of a component of a spatial satellite. The two versions of the component are described in two sets of documents. The total number of documents is 62: this number is not really high but it is justified by the fact that only a subset of documents have been selected. That is, we only select the documents that clearly identify which version of the component is concerned, i.e. the documents that concern similar goals. The main drawback of the corpus is that the number of documents is not well balanced: around one third of the documents concerns the first version of the product while the other two thirds concern the second version. However, experts of the domain are not able to select only a subset of the documents which concern the second version (i.e. what should be the selection criteria). Hence, even if we are aware of this unbalanced context, with the agreement of the experts of the domain, we proceed to the evaluation process.

Association rules have been produced with the help of Perl scripts and the reasoning stage have been made with the help of Prolog and SQL queries. We have tested several configurations for setting the values of γ (keyword threshold), σ

(support threshold) and χ (confidence support): the goal of the setting is to produce a reasonable number of rules. The configuration that has been selected is the following:

- threshold γ: the average of all positive weights;
- threshold σ: 8 documents (that is 15% of the documents relate keywords in the same way);
- threshold χ: 70% (the confidence level should be high since we consider technical documentation rather than news agency reports and thus data are highly homogeneous);

Using this setting, we obtain 1584 association rules. At first, we consider the rules that are based on documents which differ in a significant way. Among the 1584 rules, 300 of them are related to documents that do not belong to the same corpus; i.e., for each corpus, predicate `onlyOneSet` does not hold for these 300 rules. These rules represent knowledge sharing by the two corpuses. The following table details how these rules are related to documents that deeply differ.

threshold κ	number of common rules
0.1	0
0.2	29
0.25	171
0.3	241

Table 1: Different documents and common rules

Table 1 shows that threshold κ deserves fine tune otherwise too many rules are selected by predicate `commonKnowledge`. Very quickly, even if the documents differ in a deep way (threshold set at 0.3), almost 3/4 of the common rules are elicited. Let us consider the cases where $0.2 \leqslant \kappa \leqslant 0.3$; among the selected rules, let us mention the two following representative rules[1]:

$$programming \Rightarrow parameters$$
$$mode \Rightarrow parameters$$

Experts of the domain (engineers working on the design stage of the component) state that these illustrative rules, but also most of the others, do not bring a rich semantic, mainly some “basic facts”. As we will see this first limited result is mainly due to the unbalanced number of documents belonging to the two corpuses.

Let us now consider knowledge change, that is rules that have (dis)appeared even if they are linked to similar documents. Table 2 shows that very few rules appear or disappear if inter-corpus documents are (strongly) similar. We need to set a not so high threshold in order to obtain a significant number of rules. Let us consider the case where $\kappa \geqslant 0.7$ in order ti look at the content of some of the rules given by predicate `changedRules`. Among the elicited rules, let us mention the interesting results:

$$power\ supply, temperature \Rightarrow battery$$
$$battery, temperature \Rightarrow power\ supply$$

threshold κ	number of different rules
0.9 and above	3
0.7	7
0.6	33
0.5	321

Table 2: Similar Documents and different rules

These two rules are only related to the first version of the documentation. As confirmed by the experts of the domain, these rules show that the problem of energy and temperature is a more important issue in the first version of the component than in the second version.

Finally, we focus on the rules which satisfy predicate `onlyOneSet`. All these rules are only related to the second version of documentation. This high value is probably due to the unbalanced number of documents. However, the content of these rules gives an explanation about that number. Let us mention these two characteristic rules:

$$plan, review \Rightarrow documentation$$
$$requirement, quality \Rightarrow plan$$

These two rules show in a clear way that quality and documentation management is an issue of first importance during the design stage of the second version of the component. This characteristic is confirmed by the domain experts. It explains why the number of available documents is higher in the second corpus. The question is then, if we remove the documents that ”support” this quality management aspect, what results do we obtain? In order to set which documents should be no longer considered, we proceed as follows: we evaluate the relevance of the rules w.r.t. the average of the weights of the keywords used in the rule, we obtain the weight of the rule. Next we focus on the rules which are the most relevant; that is the rules with a weight above the average of the weights of all rules. With respect to this set of relevant rules, we obtain the documents which are involved in the definition of these rules and thus the documents which are the most involved in the quality management aspect. This subset of relevant documents contains 30 documents and thus the knowledge base is rebuilt w.r.t. the 32 remaining documents. The configuration for the definition of the knowledge base is the following:

- threshold γ: the average of all positive weights;
- threshold σ: 5 documents;
- threshold χ: 70%;

Using this setting, we obtain 192 association rules. All the 192 rules involved documents belonging to both versions.

threshold κ	number of common rules
0.1	6
0.2	74
0.25	151

Table 3: Different documents and common rules

Table 3 deserves more attention than in the first case since knowledge is supposed to be more homogeneous: let us

---

[1] Original documents are not written in English and in this paper we show the corresponding English terms; it also explains why, in some rules, multiple words terms are mentioned.

remark that more rules are exhibited even if the involved documents deeply differ. Let us mention the following rule ($\kappa \leqslant 0.2$):

$$interface, sending \Rightarrow message$$

which stresses the fact that in both versions one of the key issue is the transmission issue. And even if this issue in considered in several ways which may strongly differ, it can be elicited from the knowledge base. We relate this common knowledge aspect with knowledge change aspect, i.e. rules obtained with the help of predicate `changedRules`. Table 4 shows that both versions have some specific characteristics which are firmly defined. That is, the threshold for similarity does not really influence the elicitation of the rules that stress in a strong way the specificities. These rules characterize the design of the product and not the management of the design stage. The following rule illustrates this aspect:

threshold κ	number of different rules
0.9	32
0.8	34
0.6	38

Table 4: Similar Documents and different rules

$$power\ supply, temperature \Rightarrow sending$$

This rule has to be connected to the previous rules given by predicate `commonKnowledge`: it confirms the same issue (transmission) is handled in both versions but with different means or sub-issues (eg. temperature).

All these results have been presented to the experts of the domain. They agree that the results are sound and provide an overview of documentation which can be helpful. That is, a spatial satellite may have a (very) long life and it is quite common that spatial engineers rarely interact with it during long periods; meanwhile teams of engineers may change and if an intervention is required, engineers have to handle documentation and understand what were the main problems and the key issues during the design stage. In that context, our proposal is a first step towards the handling of this aspect.

Conclusion

In this paper, we have proposed a framework for reasoning about documentation change. For this, we have supposed that documentation may be partitioned in two sets and we have proposed to compare the two sets with the help of a knowledge base. The proposed knowledge base helps at stressing the difference between these two sets of documents. As illustrated by the example detailed in the previous section, we are able, with the help of the proposed framework, to focus on a quick way on changes. Let us also mention that the experts of the domain have stressed (i) the soundness of the results and (ii) the interest for these families of tools which help to manage documentation. Finally, we have seen that the proposed framework may be used to refine the definition of the different corpuses: the knowledge base helps at exhibiting biases such as the quality management aspect in the illustration; we have shown how results can be used to compose and recompose the corpuses in order to avoid biases.

As future work, the immediate milestone consists of evaluating our proposal with other documentation; this evaluation stage will help us to take care of the scale aspect (handling of (very) large documentation) and also to consider if the framework can be applied to other kinds of documents (news agency reports sorted with respect to some criteria). As a mid-term goal, we first want to consider semi-structured documents such as XML documents; using extra information such as the title or the documentation management details can be very helpful to improve the exhibition of changes. Second, we want to refine the evaluation of change by considering the evolution of associations rules, that is evaluating how the definitions of rules have changed (new consequence or antecedent).

References

Feldman, R., and Hirsh, H. 1996. Mining associations in text in the presence of background knowledge. In *Proc. of KDD'96*, 343–346.

Feldman, R., and Sanger, J. 2006. *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press.

Klemettinen, M.; Mannila, H.; Ronkainen, P.; Toivonen, H.; and Verkamo, A. I. 1994. Finding interesting rules from large sets of discovered association rules. In *Proc. of CIKM '94*, 401–407. New York, NY, USA: ACM.

Resnik, P. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proc. of IJCAI'95*, 448–453.

Richardson, R.; Smeaton, A.; and Murphy, J. 1994. Using wordnet as a knowledge base for measuring semantic similarity between words. TR CA-1294, School of Computer Applications, Dublin City University, Dublin, Ireland.

Salton, G., and Buckley, C. 1988. Term weighting approaches in automatic text retrieval. *Information Processing and Management* 24(5):513–523.

Salton, G. 1989. *Automatic text processing: the transformation, analysis, and retrieval of information by computer*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.

Silverstein, C.; Brin, S.; Motwani, R.; and Ullman, J. 2000. Scalable techniques for mining causal structures. *Data Mining and Knowledge Discovery* 4(2-3):163–192.

Tan, P.-N.; Kumar, V.; and Srivastava, J. 2002. Selecting the right interestingness measure for association patterns. In *Proc. of KDD'02*, 32–41. New York, NY, USA: ACM.

Toivonen, H.; Klemettinen, M.; Ronkainen, P.; Hätönen, K.; and Mannila, H. 1995. Pruning and grouping discovered association rules. In *Proc. of ECML'95*, 47–52.