J Intell Inf Syst (2012) 39:209–247
DOI 10.1007/s10844-011-0189-9

Towards an effective automatic query expansion process using an association rule mining approach

Chiraz Latiri · Hatem Haddad · Tarek Hamrouni

Received: 4 May 2011 / Revised: 20 November 2011 / Accepted: 24 November 2011 /
Published online: 20 December 2011
© Springer Science+Business Media, LLC 2011

Abstract The steady growth in the size of textual document collections is a key progress-driver for modern information retrieval techniques whose effectiveness and efficiency are constantly challenged. Given a user query, the number of retrieved documents can be overwhelmingly large, hampering their efficient exploitation by the user. In addition, retaining only relevant documents in a query answer is of paramount importance for an effective meeting of the user needs. In this situation, the query expansion technique offers an interesting solution for obtaining a complete answer while preserving the quality of retained documents. This mainly relies on an accurate choice of the added terms to an initial query. Interestingly enough, query expansion takes advantage of large text volumes by extracting statistical information about index terms co-occurrences and using it to make user queries better fit the real information needs. In this respect, a promising track consists in the application of data mining methods to the extraction of dependencies between terms. In this paper, we present a novel approach for mining knowledge supporting query expansion that is based on association rules. The key feature of our approach is a better trade-off between the size of the mining result and the conveyed knowledge. Thus, our association rules mining method implements results from Galois connection theory and compact representations of rules sets in order to reduce the huge number of potentially useful associations. An experimental study has examined the application of our approach to some real collections, whereby automatic query expansion has been performed. The results of the study show a significant improvement in the

C. Latiri · H. Haddad · T. Hamrouni (✉)
URPAH Team, Computer Sciences Department, Faculty of Sciences of Tunis,
El Manar University, Tunis, Tunisia
e-mail: tarek.hamrouni@fst.rnu.tn

C. Latiri
e-mail: chiraz.latiri@gnet.tn

H. Haddad
e-mail: haddad.hatem@gmail.com

performances of the information retrieval system, both in terms of recall and precision, as highlighted by the carried out significance testing using the Wilcoxon test.

Keywords Text mining · Query expansion · Information retrieval · Association rule · Generic bases · Wilcoxon test

1 Introduction and motivations

Information retrieval (IR) studies the process of determining the adequacy between a user-defined query and a document collection, usually resulting in a subset of relevant documents containing the same terms as the user query. A classical model of IR (Salton and McGill 1983) consists of assigning index terms to each document in the collection, limiting queries to the global set of index terms, and using matching measures between queries and documents seen as sets of terms. In this situation, the query expansion technique aims to reducing the usual query/document mismatch by expanding the query using terms that are related to the original query terms, but have not been explicitly mentioned in the query. The goal of this technique is not only to improve the recall by retrieving relevant documents that cannot be retrieved by the user query, but also to improve the precision of the retrieved documents by putting the most relevant ones at the top list of the retrieved documents. Correlations between terms are basically evaluated on the ground of some statistically motivated measures of terms co-occurrences within the documents of the collection (Qui and Frei 1993; Sun et al. 2006). The detection of such correlations requires the analysis of the entire document collection, or at least a large enough part of it, which is a computationally intensive task.

Interestingly enough, to address query expansion in an efficient and effective manner, we claim that a synergy between classical IR techniques and some advanced data mining methods, especially association rules (Agrawal and Skirant 1994), is particularly appropriate. This has been highlighted in some previous studies, like (Haddad et al. 2000; Lin et al. 2008; Rungsawang et al. 1999; Tangpong and Rungsawang 2000). Indeed the paradigm of association rules mining aims to the extraction of frequently occurring patterns in a transaction database that can be reasonably represented as a family of subset from a global set. In our case, the mined patterns are sets of terms (*aka* termsets[1]). A document collection can then be seen as a family of sets of terms drawn from a global set of index terms. An *association rule* is a relation $X \Rightarrow Y$, where X and Y are two sets of terms. The advantage of the insight gained through association rules is in the contextual nature of the discovered inter-term correlations. Indeed, more than a simple assessment of pair-wise term occurrences, an association rule binds two sets of terms, which respectively constitute its premise (X) and conclusion (Y) parts. Thus, a rule approximates the probability of having the terms of the conclusion in a document, given that those of the premise

[1] By analogy to the *itemset* terminology used in data mining for a set of items.

are already there. The use of such dependencies in a query expansion process should significantly increase the retrieval effectiveness since they reflect more thoroughly terms use context in the document collection.

Compared to the related works (*cf.* Section 4.1), our present study is a step towards the verification of the above hypothesis. This can be carried out through a systematic comparison of the retrieval results from the initial queries, on the one hand, and from their respective expanded versions, on the other hand. Given an initial query, the associated expanded version is obtained using the terms of conclusion parts of association rules having terms of the initial query in their premise parts.

However, applying association rules in the context of IR is far from being a trivial task, mostly because of the huge number of potentially interesting rules that can be drawn from a document collection. Moreover, detecting correlations between terms requires the analysis of the entire document collection, which is also a computationally intensive task especially whenever considering highly sized document collections. Fortunately enough, in our approach, the mined rules will constitute a knowledge source extracted *only once* and used *for each query*. The rule mining process is indeed carried out before evaluating user queries, and, thus, even if the considered document collection is large, this will not affect the running time of a given query. In this respect, we mainly concentrate in this work on reducing the number of selected rules for the expansion process while retaining the most interesting ones, and, thus eliminating the redundant ones.

Various techniques are used to limit the number of reported rules, starting by basic pruning techniques based on thresholds for both the frequency of the mined pattern (called *support*) and the dependency strength between the premise and the conclusion parts of a rule (called *confidence*) (Agrawal and Skirant 1994). More advanced techniques only producing a reduced number of the entire set of rules rely on closures and Galois connections (Bastide et al. 2000; Pasquier et al. 2005; Stumme et al. 2002), which are in turn constructs from Galois lattice theory and Formal Concept Analysis (FCA) (Ganter and Wille 1999). These compact subsets of association rules are commonly called *concise representations of association rules* or, simply, *generic bases*. These latter subsets are based on a partition of the patterns space into disjoint equivalence classes whose associated elements share the same characteristics. The maximal element into a given class is called a *closed pattern* (Pasquier et al. 2005), while the minimal elements are called *minimal generators* (Bastide et al. 2000). In recent literature, some works have yield results on these compact representations, whose impact on association rule reduction is proved (Balcázar 2010; Bastide et al. 2000; BenYahia et al. 2009; Pasquier et al. 2005; Zaki 2004).

In this paper, we propose the generation of *irredundant* association rules from a document collection based on the extraction of the *augmented Iceberg lattice*, *i.e.*, an upper set of the Galois lattice that is limited to frequent closed termsets "decorated" by the set of their minimal generators. In this context, the generators are used in the premises of the discovered rules, while the frequent closed termsets help to constitute the conclusions. The precedence relation of the Iceberg lattice allows limiting the cost of the rule extraction by avoiding some redundant combinations. Then, we detail the process using the discovered association rules in an automatic query expansion process.

1.1 Paper contributions

The contributions of this paper are summarized as follows:

1. We propose a new *minimal* generic basis, called $\mathcal{MGB}$, for only retaining irredundant association rules. The design of this basis relies on the Formal Concept Analysis (FCA) mathematical settings. In this respect, $\mathcal{MGB}$ only contains as few valid rules as possible by punning redundant ones, while each retained rule conveys a minimal premise and a maximal conclusion w.r.t. a validity criterion. Thus, each rule offers for a set of terms in its premise part the maximal possible expansion through the terms in its conclusion part. The proposed basis is then suitable for an automatic query expansion process in which the set of terms of the query will be expanded using the maximal possible set of terms located in the conclusion parts of retained rules in the basis having the terms of the query located in their premise part.
2. We present a new automatic query expansion process based on the new generic basis. The main thrust in the proposal is that the introduced basis gathers a minimal set of rules allowing an effective selection of rules to be used in the expansion process. Interestingly enough, the proposed process is generic in the sense that it is not limited to a given set of rules, these latter being used as background knowledge for the expansion process. Thus, other concise representations of association rules can be used instead of the $\mathcal{MGB}$ basis without modifying the whole expansion process.
3. To validate the proposed approach, we carry out experiments using three weighting schemes, namely $tf \times idf$ (Salton and Buckely 1988), BM25tf (Zhai 2001) and Okapi BM25 (Jones et al. 2000). We then apply our expansion method to the OFIL and INIST document collections of the second Amaryllis campaign,[2] as well as the Le Monde 94 and ATS 94 document collections of the CLEF 2003 campaign (Collection 2001).[3] A fifth collection composed of both Le Monde 94 and ATS 94 documents is also tested. In our experiments, we use the classical performance criteria of *recall* and *precision* (Salton and McGill 1983).

It is worth noting that it is out of the scope of this work to discuss how the generic basis is efficiently discovered. Indeed, mainly the design of this latter is thoroughly discussed in this work. This is argued by two facts. On the one hand, as aforementioned, the mining of irredundant association rules is carried out before starting the query process. Thus, it does not affect the performance of this process. In this respect, the proposed basis, through the form of the rules it contains, makes it possible the optimization of the expansion process since as detailed in the remainder only few rules are retained. On the other hand, the efficient mining of the proposed basis can benefit from several existing algorithms in the literature dedicated to frequent patterns mining from large databases of millions of transactions (El-Hajj and Zaiane 2005; Lucchese et al. 2003).

[2]The Amaryllis project is initiated by INIST-CNRS and co-funded by AUPELF-UREF. Its goal is to evaluate French Text retrieval systems.

[3]The Cross-Language Evaluation Forum (CLEF) promotes multilingual information access. It offers benchmark collection data for evaluating IR systems. The associated website is: http://www.clef-campaign.org/.

1.2 Paper organization

The remainder of the paper is organized as follows: Section 2 recalls the basic mathematical foundations for the derivation of association rules based on Iceberg lattice. We present in Section 3 an overview of the literature dedicated to the extraction of generic bases. We also introduce a novel minimal generic basis of irredundant association rules. The proposed basis is then compared theoretically and experimentally with those of the literature. Section 4 discusses related works on query expansion for information retrieval. Then, a detailed description of our approach for query expansion based on irredundant association rules is presented. Section 5 describes the results of the experiments carried out on five document collections. The conclusion and future work are finally presented in Section 6.

2 Association rules mining based on iceberg lattice

After introducing some notations, we state the formal definitions of the concepts used in the remainder of the paper. In this respect, Table 1 provides an overview of the notations used in this and later sections.

In this paper, we shall use in text mining field, the theoretical framework of Formal Concept Analysis (FCA) presented in Ganter and Wille (1999). First, we formalize an extraction context made up of documents and index terms, called *textual context*.

Definition 1 A *textual context* is a triplet $\mathfrak{M} := (\mathcal{C}, \mathcal{T}, \mathcal{I})$ where:

- $\mathcal{C} := \{d_1, d_2, \ldots, d_n\}$ is a finite set of n documents of a collection.
- $\mathcal{T} := \{t_1, t_2, \cdots, t_m\}$ is a finite set of m distinct terms in the collection. The set $\mathcal{T}$ then gathers without duplication the terms of the different documents which constitute the collection.
- $\mathcal{I} \subseteq \mathcal{C} \times \mathcal{T}$ is a binary (incidence) relation. Each couple $(d, t) \in \mathcal{I}$ indicates that the document $d \in \mathcal{C}$ has the term $t \in \mathcal{T}$.

Example 1 Consider the context given in Table 2, used as a running example through this paper. Here, $\mathcal{C} := \{d_1, d_2, d_3, d_4, d_5, d_6\}$ and $\mathcal{T} := \{\text{A}, \text{C}, \text{D}, \text{T}, \text{W}\}$. The couple $(d_2, \text{C}) \in \mathcal{I}$ since it is crossed in the matrix. This denotes that the document d_2 contains the term C. On the contrary, since the term W does not appear in the document d_6, the associated cell to the couple (d_6, W) is not crossed in the matrix. Thus, $(d_6, \text{W}) \notin \mathcal{I}$.

Table 1 Summary of notations

Notation	Description
$\mathcal{C}$	the ***whole set*** of documents which form the collection
D	a ***set*** of documents belonging to the collection ($D \subseteq \mathcal{C}$)
d	a ***single*** document of the collection ($d \in \mathcal{C}$)
$\mathcal{T}$	the ***whole set*** of ***distinct*** terms of the collection $\mathcal{C}$
T	a ***set*** of terms of the collection ($T \subseteq \mathcal{T}$)
t	a ***single*** term of the collection ($t \in \mathcal{T}$)

Table 2 A textual context $\mathfrak{M} := (\mathcal{C}, \mathcal{T}, \mathcal{I})$

$\mathcal{I}$	A	C	D	T	W
d_1	×	×		×	×
d_2		×	×		×
d_3	×	×		×	×
d_4	×	×	×		×
d_5	×	×	×	×	×
d_6		×	×	×	

A termset is a set of terms. For example, {A, C, W} is a termset composed by the terms A, C and W. In the remainder, we use a separator-free form for the sets, *e.g.*, ACW stands for the termset {A, C, W}. The support of a termset is defined as follows.

Definition 2 Let $T \subseteq \mathcal{T}$. The support of T in $\mathfrak{M}$ is equal to the number of documents in $\mathcal{C}$ containing all the term of T. The support is formally defined as follows:[4]

$$Supp(T) = |\{d | d \in \mathcal{C} \ \wedge \ \forall\, t \in T : (d,\ t) \in \mathcal{I}\}| \tag{1}$$

Supp(T) is called the *absolute* support of T in $\mathfrak{M}$. The *relative* support (*aka* frequency) of $T \in \mathfrak{M}$ is equal to $\frac{Supp(T)}{|\mathcal{C}|}$.

A termset is said *frequent* (*aka large* or *covering*) if its terms co-occur in the collection a number of times greater than or equal to a user-defined support threshold, denoted *minsupp*. Otherwise, it is said *unfrequent* (*aka rare*).

Example 2 Consider the context given in Table 2 and the *minsupp* threshold set to 3. The termset AC is frequent since $Supp(\text{AC}) = 4 \geq 3$. On the contrary, the termset CDT is unfrequent since $Supp(\text{CDT}) = 2 < 3$.

Now, we establish relationships between the absolute and relative support measures and the standard ones used in IR namely the term frequency, denoted tf, and the document frequency, denoted df, whose formulae are as follows. Let d be a document of $\mathcal{C}$, and t be a term of $\mathcal{T}$:

$$tf(t, d) = \frac{|terms(d)|}{|d|} \tag{2}$$

where $terms(d)$ denotes the number of occurrences of t in d, while $|d|$ denotes the total number of terms in d.

$$df(t, \mathcal{C}) = \frac{|documents(t)|}{|\mathcal{C}|} \tag{3}$$

where $documents(t)$ denotes the number of documents containing at least an occurrence of t, while $|\mathcal{C}|$ denotes the number of documents in $\mathcal{C}$.

The term frequency indicates how many times a term appeared in a given document. On the other hand, the document frequency shed lights on the appearance degree of t within the different documents of the collection. Thus, both measures are

[4] In this paper, we denote by $|X|$ the cardinality of the set X.

dedicated to a single term. The support measure is used to indicate how many times a set of terms - a termset - simultaneously appears in the documents of the collection.

If we consider the termset T as being reduced to a single term, the relative support corresponds to the document frequency measure, w.r.t. the standard IR terminology. Indeed, $Supp(T)$ is equal to the number of documents containing T, while $|\mathcal{C}|$ is equal to the number of all documents. On the other hand, in our work, we are mainly interested in improving the expansion process. As a consequence, we only consider distinct terms within each document, thus omitting the duplication of a term in a given document. The classical measure of term frequency used in the IR terminology can then not be considered here, even for a single term.

The derivation of association rules between terms is achieved starting from the set of frequent termsets extracted from a context $\mathfrak{M}$. Many representations of frequent termsets were proposed in the literature where terms are characterized by the frequency of their co-occurrence. The ones based on *closed termsets* (Pasquier et al. 2005) and *minimal generators* (Bastide et al. 2000) are at the core of the definitions of almost all generic bases of the literature. They result from the mathematical background of FCA (Ganter and Wille 1999), described in the next subsection.

2.1 Mathematical foundations: key FCA settings

In the following, we recall basic definitions of the Galois lattice-based paradigm in FCA (Ganter and Wille 1999) and its applications to association rules mining.

2.1.1 Galois closure operator

Two functions are defined in order to map sets of documents to sets of terms and *vice versa*. Thus, for $T \subseteq \mathcal{T}$, we define:

$$\Psi(T) := \{d | d \in \mathcal{C} \ \wedge \ \forall\, t \in T : (d,\ t) \in \mathcal{I}\} \tag{4}$$

$\Psi(T)$ is equal to the set of documents containing all the terms of T. Its cardinality is then equal to $Supp(T)$.

For a set $D \subseteq \mathcal{C}$, we define:

$$\Phi(D) := \{t | t \in \mathcal{T} \ \wedge \ \forall\, d \in D : (d,\ t) \in \mathcal{I}\} \tag{5}$$

$\Phi(D)$ is equal to the set of terms appearing in all the documents of D.

Both functions Ψ and Φ constitute *Galois operators* between the sets $\mathcal{P}(\mathcal{T})$ and $\mathcal{P}(\mathcal{C})$. Consequently, the compound operator $\Omega := \Phi \circ \Psi$ is a *Galois closure operator* which associates to a termset T the whole set of terms which appear in *all* documents where the terms of T co-occur. This set of terms is equal to $\Omega(T)$. In fact, $\Omega(T) = \Phi \circ \Psi(T) = \Phi(\Psi\ (T))$. If $\Psi\ (T) = D$, then $\Omega(T) = \Phi(D)$.

Example 3 Consider the context given in Table 2. Since both terms A and C simultaneously appear in the documents d_1, d_3, d_4, and d_5, we have: $\Psi(\text{AC}) = \{d_1, d_3, d_4, d_5\}$. On the other hand, since the documents d_1, d_3, d_4, and d_5 share the terms A, C, and W, we have: $\Phi(\{d_1, d_3, d_4, d_5\}) = \text{ACW}$. It results that $\Omega(\text{AC}) = \Phi \circ \Psi(\text{AC}) = \Phi(\Psi(\text{AC})) = \Phi(\{d_1, d_3, d_4, d_5\}) = \text{ACW}$. Thus, $\Omega(\text{AC}) = \text{ACW}$. In other words, the term W appears in all documents where A and C co-occur.

2.1.2 Frequent closed termset

A termset $T \subseteq \mathcal{T}$ is said to be *closed* if $\Omega(T) = T$. A closed termset is then the maximal set of terms common to a given set of document. A closed termset is said to be *frequent* w.r.t. the *minsupp* threshold if $Supp(T) = |\Psi(T)| \geq minsupp$ (Pasquier et al. 2005). Hereafter, we denote by FCT a frequent closed termset.

Example 4 With respect to the previous example, ACW is a closed termset since there is not another term appearing in all documents containing ACW. ACW is then the maximal set of terms common to the documents $\{d_1, d_3, d_4, d_5\}$. We then have: $\Omega(\text{ACW}) = \text{ACW}$. If *minsupp* is set to 3, ACW is also frequent since $|\Psi(\text{ACW})| = |\{d_1, d_3, d_4, d_5\}| = 4 \geq 3$.

The next property states the relation between the support of a termset and that of its closure.

Property 1 The support of a termset T is equal to the support of its closure $\Omega(T)$, which is the smallest FCT containing T, *i.e.*, $Supp(T) = Supp(\Omega(T))$ (Bastide et al. 2000).

Example 5 Since $\Omega(\text{AC}) = \text{ACW}$, we have: $Supp(\text{AC}) = Supp(\text{ACW}) = 4$.

2.1.3 Minimal generator

A termset $g \subseteq \mathcal{T}$ is a *minimal generator* of a closed termset T, if and only if $\Omega(g) = T$ and $\nexists\, g' \subset g$: $\Omega(g') = T$ (Bastide et al. 2000).

Example 6 The termset DW is a minimal generator of CDW since $\Omega(\text{DW}) = \text{CDW}$ and none of its proper subsets has CDW for closure. Indeed, $\Omega(\text{D}) = \text{CD}$ and $\Omega(\text{W}) = \text{CW}$.

Corollary 1 *Let g be a minimal generator of a frequent closed termset T. According to Property* 1*, the support of g is equal to the support of its closure, i.e., $Supp(g) = Supp(T)$.*

Thus, a FCT occurs within the same set of documents and, hence, have the same support as its generators. A FCT then represents a maximal terms group sharing the same documents, while its minimal generators are the smallest incomparable elements describing the documents set. A closed termset then includes the most specific expression describing the associated documents, while a minimal generator includes one of the most general expressions. In the remainder, for each frequent closed termset T, we denote by $\mathcal{G}_T$ the set of its minimal generators.

2.1.4 Iceberg lattice

Let $\mathcal{FCT}$ be the set of frequent closed termsets of a given context. When the set $\mathcal{FCT}$ is partially ordered w.r.t. set inclusion, the resulting structure only preserves the *Join* operator (Ganter and Wille 1999). This structure is called a *join semi-lattice* or an *upper semi-lattice*; and is hereafter referred to as *Iceberg lattice* (Stumme et al. 2002).

Table 3 The set $\mathcal{FCT}$ of frequent closed termsets associated to their respective minimal generators and supports

Minimal generator(s)	FCT	Support
C	C	6
W	CW	5
D	CD	4
T	CT	4
A	ACW	4
AT/TW	ACTW	3
DW	CDW	3

In this respect, we present in this paper an approach that relies on irredundant association rules mining starting from the *augmented Iceberg lattice*, denoted by $\mathcal{AL} := (\mathcal{FCT}, \subseteq)$, which is the standard Iceberg lattice where each FCT is associated to its minimal generators.

Example 7 Consider the context given in Table 2. The *minsupp* threshold is set to 3. Table 3 shows for each frequent closed termsets, the corresponding minimal generators, and support. The associated augmented Iceberg lattice is depicted in Fig. 1, in which the minimal generators associated to each FCT are given between brackets.

Each frequent closed termset T in the Iceberg lattice has an *upper cover* which consists of the closed termsets that immediately cover T in the Iceberg lattice. This set is formally defined as follows:

$$Cov^u(T) := \{T_1 \in \mathcal{FCT} \mid T \subset T_1 \text{ and } \nexists\, T_2 \in \mathcal{FCT}\colon T \subset T_2 \subset T_1\}$$

Clearly for an element T without any proper superset in the lattice, its upper cover set is empty.

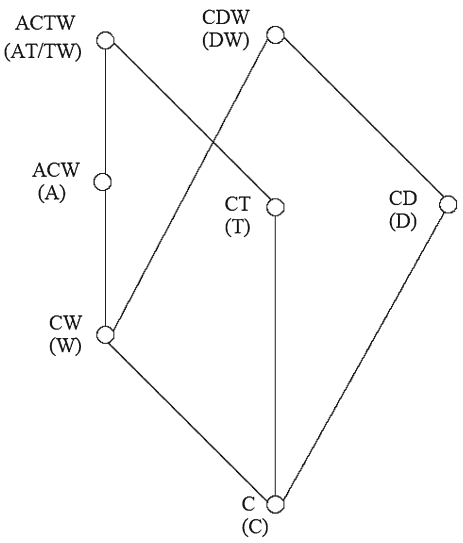

Fig. 1 The augmented Iceberg lattice

Example 8 Let us consider the FCT CW of the Iceberg lattice depicted by Fig. 1. Then, we have: $Cov^u(\text{CW}) = \{\text{ACW}, \text{CDW}\}$.

2.2 Mining association rules between terms

An association rule R is an implication of the form R: $T_1 \Rightarrow T_2$, where T_1 and T_2 are subsets of $\mathcal{T}$, and $T_1 \cap T_2 = \emptyset$. The termsets T_1 and T_2 are, respectively, called the *premise* and the *conclusion* of R. The rule R is said to be based on the termset T equal to $T_1 \cup T_2$. The *support* of a rule R: $T_1 \Rightarrow T_2$ is then defined as:

$$Supp(R) = Supp(T), \tag{6}$$

while its *confidence* is computed as:

$$Conf(R) = \frac{Supp(T)}{Supp(T_1)}. \tag{7}$$

An association R is said to be *valid* if its confidence value, *i.e.*, $Conf(R)$, is greater than or equal to a user-defined threshold denoted *minconf*.[5] This confidence threshold is used to exclude non valid rules. Also, the given support threshold *minsupp* is used to remove rules based on termsets T that do not occur often enough, *i.e.*, rules having $Supp(T) < minsupp$.

Example 9 Starting from the context depicted in Table 2, the association rule R: W $\Rightarrow$ CD can be derived. In this case, $Supp(\text{R}) = Supp(\text{CDW}) = 3$, while $Conf(\text{R}) = \frac{Supp(\text{CDW})}{Supp(\text{W})} = \frac{3}{5}$. If we consider the *minsupp* and *minconf* thresholds respectively equal to 3 and 0.5, the considered rule R is valid since $Supp(\text{R}) = 3 \geq 3$ and $Conf(\text{R}) = \frac{3}{5} = 0.6 \geq 0.5$.

In the remainder of the paper, we distinguish two types of association rules: *exact association rules* (with confidence equal to 1) and *approximate association rules* (with confidence less than 1) (Zaki 2004). This distinction is due to the fact that exact association rules and approximate ones have different properties, as shown by the following two properties.

Property 2 Approximate association rules of the form $T_1 \Rightarrow (T_2 - T_1)$ are implications between two frequent termsets T_1 and T_2 such that $T_1 \subseteq T_2$ and the closure of T_1 is a subset of the closure of T_2, *i.e.*, $\Omega(T_1) \subset \Omega(T_2)$ (Zaki 2004).

Example 10 The rule W $\Rightarrow$ (CDW—W) is an approximate one since $\Omega(\text{W}) = \text{CW} \subset \Omega(\text{CDW}) = \text{CDW}$. In this regard, its confidence is equal to $\frac{3}{5}$.

Property 3 Exact association rules $T1 \Rightarrow (T_2 - T_1)$ are implications between two frequent termsets T_1 and T_2, such that $T_1 \subseteq T_2$ and T_1 and T_2 have identical closures, *i.e.*, $\Omega(T_1) = \Omega(T_2)$ (Zaki 2004).

[5] In the remainder, $T_1 \overset{c}{\Rightarrow} T_2$ indicates that the rule $T_1 \Rightarrow T_2$ has a value of confidence equal to c.

Example 11 The rule DW $\Rightarrow$ (CDW—DW) is an exact one since Ω (DW) = Ω(CDW) = CDW. In this regard, its confidence is equal to 1.

Given a document collection, the problem of mining association rules between terms consists in generating all association rules given user-defined *minsupp* and *minconf* thresholds. This problem can be split into two steps as follows:

1. Extract all frequent termsets that occur in the collection with a support value greater than or equal to *minsupp*: The last decade witnessed many research works in order to develop efficient algorithms for mining frequent patterns. To avoid prohibitive frequent termsets number, many researchers focused on closed termsets paradigm. In fact, the set $\mathcal{FCT}$ of frequent closed termsets is usually much smaller than the set of all frequent termsets. Please refer to Ben Yahia et al. (2006) for a theoretical and an experimental comparative study of mining algorithms for $\mathcal{FCT}$.
2. Generate valid association rules (*aka* Apriori rules (Agrawal and Skirant 1994)) between terms from frequent termsets;*i.e.*, rules whose confidence are greater than or equal to *minconf*: These rules can be generated in a straightforward manner, *i.e.*, without any further access to the context (Agrawal and Skirant 1994). Nevertheless, the number of discovered association rules may grow up to several millions (Zaki 2004) while a large number of them could be redundant (Ashrafi et al. 2007; Balcázar 2010; BenYahia et al. 2009).

However, the problem of mining association rules between terms deals with a real challenge: how to only retrieve the most relevant associations from the huge number of possibilities ($2^{|T|} - 1$ for a frequent termset T)?[6]

Several approaches in the literature deal with the redundancy problem. For instance, some works relied on the use of other quality measures in the mining step in addition to the support and confidence, like lift and conviction (Guillet and Hamilton 2007). Other approaches introduced user-defined constraints during the mining process or on a post-processing step (Liu et al. 2009). More advanced techniques that produce only a limited number of rules rely on Galois closure (Ganter and Wille 1999). These techniques focus on extracting irreducible nuclei of all association rules, called *generic basis*, from which the remaining association rules can be derived (Balcázar 2010; Bastide et al. 2000; BenYahia et al. 2009; Kryszkiewicz 2002; Zaki 2004).

In this respect, the derivation of association rules is then achieved in our approach starting from the set $\mathcal{FCT}$ of frequent closed termsets extracted from the textual context in order to only retain a compact set of rules and, then, remove as much as possible the redundancy arising within association rules. In the next subsection, we present a brief review of the most used generic bases in the literature. Then, we introduce a new generic basis whose rules will be used in our automatic query expansion.

[6]The rule $T \Rightarrow \emptyset$ is usually considered as not informative.

3 Generic basis of association rules

In what follows, we focus on the results issued from FCA, to retrieve a reduced set of rules, *i.e.*, *generic basis* (Balcázar 2010; Bastide et al. 2000; BenYahia et al. 2009; Kryszkiewicz 2002). Almost all generic bases convey association rules presenting implications between minimal generators and closed termsets ensuring obtaining association rules with minimal premise and maximal conclusion part. Such rules convey the maximum of information, and are hence qualified as the most informative ones (Bastide et al. 2000). Indeed, a generic basis has to fulfill the following requirements (Kryszkiewicz 2002):

- *Informativity*: The generic basis of association rules allows one to retrieve *exactly* the support and confidence of the derived (redundant) association rules.
- *Derivability*: An inference mechanism should be provided (*e.g.*, an axiomatic system). The axiomatic system has to be valid (*i.e.*, should forbid derivation of non valid association rules) and complete (*i.e.*, should enable derivation of *all* valid association rules).
- *Compacity*: Mining the *most reduced* set of generic rules allowing the derivation of all confident remaining ones, *i.e.*, redundant rules.

In the literature, two main classes of approaches have been explored for extracting generic bases. The first class contains those which offer bases with information loss, *i.e.*, they do not fulfill the derivability or the informativity condition, while the second class covers those leading to generic bases without information loss. An interesting discussion about the main generic bases of association rules is proposed in BenYahia et al. (2009).

In the following, we firstly present the main representative generic basis of each one of the aforementioned classes. Note that the associated definitions are adapted to our context, *i.e.*, text mining, through the use of *termsets* instead of *itemsets* (*i.e.*, sets of items) initially adopted in the original works presenting these bases. Then, we introduce a new generic basis for association rules that will be used in the proposed automatic query expansion process. This section ends with a theoretical and an experimental comparison of the proposed generic basis with those of the literature.

3.1 Extraction of generic bases without information loss

In the literature, many approaches were interested in reducing the set of mined association rules without any loss of information (*cf.* for example Balcázar 2010, BenYahia et al. 2009). However, as indicated in BenYahia et al. (2009), Kryszkiewicz (2002), the main representative of this class of generic bases is that of Bastide et al. (2000). In their work, the authors considered the following rule-redundancy definition:

Definition 3 Let $\mathcal{VAR}$ be the set of valid association rules that can be drawn from a textual context for a minimum support threshold *minsupp* and a minimum confidence threshold *minconf*. An association rule R_1: $T_1 \Rightarrow T_2 \in \mathcal{VAR}$ is said *redundant with respect to* (or *derivable from*) a rule R_2: $T'_1 \Rightarrow T'_2 \in \mathcal{VAR}$ *iff*:

1. $Supp(R_1) = Supp(R_2)$ and $Conf(R_1) = Conf(R_2)$, and,
2. $T'_1 \subseteq T_1$ and $T_2 \subset T'_2$.

Example 12 Consider both rules R_1: W $\Rightarrow$ A and R_2: W $\Rightarrow$ AC. Since, Ω(AW) = Ω(ACW) = ACW, we have based on Property 1, $Supp$(AW) = $Supp$(ACW). Thus, $Supp(R_1) = Supp(R_2)$. Moreover, since both rules share the same premise, they then have the same confidence: $Conf(R_1) = Conf(R_2)$. Consequently, R_1 is redundant w.r.t. R_2 since they share the same support and confidence values as well as the same premise while the conclusion of R_1, namely A, is a proper subset of that of R_2, namely AC.

Based on Definition 3, for an association rule $T_1 \Rightarrow T_2$, if there is not another rule $T_1' \Rightarrow T_2'$ such that $Supp(R_1) = Supp(R_2)$, $Conf(R_1) = Conf(R_2)$, $T_1' \subseteq T_1$, and $T_2 \subset T_2'$, then $T_1 \Rightarrow T_2$ is said *minimal non-redundant* (Bastide et al. 2000). Note that this definition ensures that non-redundant association rules will hence have *minimal premises* and *maximal conclusions*. The authors they characterize what they called "the generic basis for exact association rules", denoted $\mathcal{GBE}$, and the "informative basis for approximate association rules", denoted $\mathcal{GBA}$. The $\mathcal{GBE}$ and $\mathcal{GBA}$ bases are defined as follows (Bastide et al. 2000):

Definition 4 Let $\mathcal{FCT}$ be the set of frequent closed termsets extracted from a textual context and, for each frequent closed termset T, $\mathcal{G}_T$ denotes the set of minimal generators of T. The generic basis for exact association rules is defined as follows:

$$\mathcal{GBE} := \{R : g \Rightarrow (T - g) |\ T \in \mathcal{FCT} \ \wedge\ g \in \mathcal{G}_T \ \wedge\ g \neq T\}. \tag{8}$$

The generic basis for approximate association rules is defined as follows:

$$\mathcal{GBA} := \big\{R : g \Rightarrow (T - g) |\ T, T_1 \in \mathcal{FCT} \ \wedge\ g \in \mathcal{G}_{T_1} \ \wedge\ T_1 \subset T \ \wedge\ Conf(R) \geq minconf\big\}. \tag{9}$$

With respect to Definition 4, we consider that given an Iceberg lattice, representing precedence-relation within closed termsets, generic basis of association rules can be derived in a straightforward manner. We assume that in such structure, each closed termset is augmented with its associated list of minimal generators. Hence, approximate rules (ARs) represent *inter-node* implications, assorted with the confidence, from a sub-closed-termset to a super-closed-termset while starting from a given node in an ordered structure. On the other hand, exact rules (ERs) are *intra-node* implications extracted from each node in the ordered structure.

Example 13 We refer in this example to the augmented Iceberg lattice depicted by Fig. 1 and Table 3. Consider for example the frequent closed termset ACTW and its minimal generator AT. The induced rule based on these patterns is: AT $\Rightarrow$ CW which belongs to $\mathcal{GBE}$.

On the other hand, suppose $minconf = 0.5$ and consider both FCTs CW and ACTW. Since CW $\subset$ ACTW and W is a minimal generator of CW, the rule W $\Rightarrow$ ACT belongs to $\mathcal{GBA}$ and has a confidence equal $\frac{3}{5} \geq minconf$.

To overcome the weaknesses resulting from an oversized $\mathcal{GBA}$ basis and its low compactness rate especially for sparse contexts, Bastide et al. (2000) defined the transitive reduction of the generic basis of the approximate generic rules, denoted $\mathcal{TGBA}$, as follows:

Definition 5 The $\mathcal{TGBA}$ basis is equal to:

$$\mathcal{TGBA} := \{R : g \Rightarrow (T - g) |\ T, T_1 \in \mathcal{FCT} \ \wedge \ T \in Cov^u(T_1) \wedge g \in \mathcal{G}_{T_1} \ \wedge \ Conf(R) \geq minconf\}. \quad (10)$$

Example 14 Suppose $minconf = 0.5$. Since ACW $\in Cov^u$(CW) and W is a minimal generator of CW, the rule W $\Rightarrow$ AC belongs to $\mathcal{TGBA}$ and has a confidence equal $\frac{4}{5} \geq minconf$. Note, however, that the rule W $\Rightarrow$ ACT belonging to $\mathcal{GBA}$ (*cf.* Example 13) is not considered in $\mathcal{TGBA}$ since ACTW $\notin Cov^u$(CW).

In Kryszkiewicz (2002), the author proved that the couple $(\mathcal{GBE}, \mathcal{GBA})$ forms a sound and informative generic basis, *i.e.*, the respective support and confidence of inferred rules can be exactly retrieved. However, as a drawback, such generic basis may be over-sized, especially for dense contexts.

3.2 Extraction of generic bases with information loss

Some generic bases, with information loss, were proposed in the literature. The main representative of this class is that introduced in Zaki (2004). In this work, Zaki introduced a generic basis called the *non-redundant association rule basis*, denoted $\mathcal{NRR}$. He presented an approach based on an axiomatic system, taking into account support and confidence, for the generation of the whole set of association rules from a minimal rule basis. The author considered the following rule-redundancy definition:

Definition 6 Let $\mathcal{VAR} := \{R_1, R_2, \ldots, R_n\}$ be the set of valid association rules that may be drawn from a textual context $\mathfrak{M}$. R_1: $T_1 \Rightarrow T_2 \in \mathcal{VAR}$ is subsuming the rule R_2: $T_1' \Rightarrow T_2'$ (or equivalently, R_2 is redundant w.r.t. R_1), denoted by $R_1 \preceq R_2$, if and only if the following conditions are fulfilled:

1. $T_1 \subseteq T_1'$ and $T_2 \subseteq T_2'$;
2. $Supp(R_1) = Supp(R_2)$ and $Conf(R_1) = Conf(R_2)$.

Consequently, an association rule R_2 is considered as redundant *iff* it exists an association rule R_1 such that $R_1 \preceq R_2$, otherwise it is said to be non-redundant.

Example 15 Consider the augmented Iceberg lattice depicted by Fig. 1 and Table 3. With respect to Definition 6, the rule R_1: W $\Rightarrow$ T subsumes the rule R_2: CW $\Rightarrow$ AT since, on the one hand, W $\subseteq$ CW and T $\subseteq$ AT and, on the other hand, $Supp(R_1) = Supp(R_2) = 3$ and $Conf(R_1) = Conf(R_2) = \frac{3}{5}$. The rule R_2 is then considered redundant w.r.t. R_1.

The notion of non-redundancy considered by Zaki is related to the inference system composed of the transitivity axiom of Luxenburger (1991) and that of the augmentation of Armstrong (1974). Hence, all the generated rules have minimal premise and minimal conclusion parts. However, as we will show later in this paper,

this minimal form does not always catch all implicit knowledge hidden in the document collection.

Based on Definition 6, Zaki introduced the $\mathcal{NRR}$ basis as follows:

Definition 7 Let $\mathcal{VAR} := \{R_1, R_2, \ldots, R_n\}$ be the set of valid association rules that may be extracted from a context $\mathfrak{M}$. Thus,

$$\mathcal{NRR} := \{R_i \in \mathcal{VAR} \mid \nexists R_j \in \mathcal{VAR} : \ R_j \preceq R_i \ \wedge \ i \neq j\}. \tag{11}$$

Example 16 For $minsupp = 3$ and $minconf = 0.5$, the rule $W \Rightarrow T \in \mathcal{NRR}$ since this rule is valid and there is not another rule which subsumes it.

Nevertheless, as pointed in BenYahia et al. (2009), the $\mathcal{NRR}$ basis does not cover all the valid rule set. Some valid rules do not belong to the $\mathcal{NRR}$ basis and are not derivable by the proposed axiomatic system. Moreover, w.r.t. Definition 6, if an association rule R_2 is considered as redundant w.r.t. R_1, then it should inherit the same support and confidence values of R_1. Nevertheless, an association rule R_3, inferred by applying the Luxenburger transitivity axiom on R_1 and R_2 from the $\mathcal{NRR}$ basis given in Definition 6, may have a confidence value which is different from those of both R_1 and R_2.

It is worth recalling that applying association rules in the context of IR is far from being an easy task, mostly because of the huge number of potentially interesting rules that can be drawn from a document collection. In order to dramatically reduce the large number of rules, we introduce in the next subsection a new approach for mining a *minimal generic basis* of irredundant association rules from text dealing with this task.

3.3 $\mathcal{MGB}$: a minimal generic basis of irredundant association rules between terms

In an information retrieval process, the use of association rules for query expansion is carried out as follows: if the premise part T_1 of a valid rule R: $T_1 \Rightarrow T_2$ is contained in the initial query, the terms in the conclusion part T_2 are used to extend those of the query. Indeed, the presence in the corpus of the set of terms that constitutes T_1 implies that of a distinct set of terms that forms T_2 with a certain probability (conveyed through the confidence measure).

However, the main challenge when using association rules in an automatic query expansion consists in the large number of rules that can be applied for expanding an initial query. Indeed, each valid rule whose premise is contained in the query can be used to extend it. On the other hand, the expansion process relies on an inclusion test to check whether the premise part terms of a candidate rule for the expansion appear in the query to be expanded or not.

In this situation, we propose the construction of a new *minimal generic basis*, called $\mathcal{MGB}$, based on the extraction of the *augmented Iceberg lattice*. In this context, the generators are used in the premise part of the discovered rules, while the closed termsets help to constitute the conclusions. The precedence relation of the augmented Iceberg lattice allows limiting the cost of the rules extraction by avoiding some redundant combinations. Although the proposed $\mathcal{MGB}$ generic basis can be

used in several application fields, it will be shown later in details that it is suitable for automatic query expansion based on association rules. This is carried out thanks to its design taking into account the following important features:

1. The $\mathcal{MGB}$ basis offers a reduced size w.r.t. the whole set of valid association rules and even to the other generic basis proposed in the literature. This is an important point since it avoids as much as possible the redundancy amongst association rules while retaining as few valid rules as possible. The expansion process based on $\mathcal{MGB}$ rules then only uses a limited number of rules. This is important since avoiding the cost of a combinatory process if a large number of rules is used. Indeed, although automatic, the query expansion process is launched once the initial query given and, then, optimizing running time of this process is important for user satisfaction.
2. The rules conveyed in $\mathcal{MGB}$ have an interesting form. Indeed, each retained rule conveys a minimal premise, *i.e.*, containing a minimal generator, which implies the maximal possible conclusion w.r.t. the validity criterion through *minconf*. The rule is then based on a closed termset which is the largest set of terms whose presence in the document collection depends on that of the set of terms of the premise with a probability greater than or equal to *minconf*. Such a rule form has three complementary advantages: (i) its premise part being minimal for each rule allows reducing the cost of existence test of the terms of the premise in a query to be expanded, (ii) its maximal conclusion offers different possibilities to expand a given query only using the terms of the premise, and (iii) only allowing the largest possible conclusion, while fulfilling the validity property, dramatically reduces the number of retained rules in the basis which optimizes further use as for query expansion.
3. As shown later, although its design has for a main purpose the compactness aspect, for a given *minsupp* and a given *minconf*, the $\mathcal{MGB}$ basis allows deriving all valid association rules without information loss. Thus, $\mathcal{MGB}$ also derives rules which constitute the other generic bases. The terms resulting from the expansion of a query that may offer another basis are then necessarily obtained using $\mathcal{MGB}$. Interestingly enough, relying on $\mathcal{MGB}$ results in smaller-size storage requirement for later use, while making easier further manipulations. Interestingly enough, the automatic query expansion approach we will propose is not limited to only handle generic rules belonging to $\mathcal{MGB}$. Indeed, it allows the use of any set of association rules as background knowledge, in particular those which constitute another generic basis (Balcázar 2010; Bastide et al. 2000; BenYahia et al. 2009; Zaki 2004).

The main features of $\mathcal{MGB}$ will be further detailed in the following paragraphs.

3.3.1 Irredundant association rules discovery

An association rule R_2 is said to be redundant w.r.t. a rule R_1 if the information conveyed by R_1 implies the information conveyed by R_2.

We believe that w.r.t. the specific objectives of query expansion, the redundancy definition proposed by Zaki (2004) is inadequate since it results in a set of rules with minimal conclusion parts and therefore an additional effort is necessary to achieve full-scale expansion. The definition of Bastide et al. (2000), in turn, suffers from high permissiveness. Indeed, having exactly the same confidence as elimination criteria

fails to exclude rules of identical premises and comparable conclusions w.r.t. set inclusion.

In this paper, we consider rules that maximize the number of terms in the conclusion. The idea behind this is to obtain additional relevant documents through expanded queries. So, we define redundancy as follows:

Definition 8 An association rule R_1: $T_1 \Rightarrow T_2$ is redundant w.r.t. a rule R_2: $T'_1 \Rightarrow T'_2$ if and only if one of the following conditions is fulfilled:

1. $\Omega(T'_1 \cup T'_2)=\Omega(T_1 \cup T_2)$ and $T'_1 \subseteq T_1$,
2. $T'_1 = T_1$ and $T_2 \subset T'_2$.

Example 17 Consider a frequent closed termset ACTW, and the two rules R_1: A $\Rightarrow$ CTW and R_2: AC $\Rightarrow$ TW. With respect to the above mentioned definition, R_2 is considered redundant w.r.t. R_1 since the discovery of R_1 implies necessarily that of R_2. Indeed, the two rules have the same support equal to $Supp$(ACTW) and their respective confidences are $Conf(R_1) = \frac{Supp(\text{ACTW})}{Supp(\text{A})}$ and $Conf(R_2) = \frac{Supp(\text{ACTW})}{Supp(\text{AC})}$. We deduce that $Conf(R_2) \geq Conf(R_1)$. Thus, if R_1 is a valid rule, necessarily will also be R_2. In addition, R_1 is more suitable than R_2 for an expansion process since it offers several candidates for the expansion through the terms in conclusion, among which some do not appear in the conclusion of R_2.

We now define the valid premises w.r.t. the generic basis we will propose. In this respect, we begin by defining the set of all potential ones.

Definition 9 Let T be a frequent closed termset (FCT). The set of all potential premises of valid rules based on T contains its minimal generators as well as those of FCTs included in T. It is defined as follows:

$$allG_T := \{g \subseteq \mathcal{T} | \Omega(g) = T_1 \subseteq T\}. \tag{12}$$

Example 18 Consider the FCT ACW: $allG_{ACW}$ = {A, C, W}.

Since in our approach, we aim to only retaining valid rules with minimal premises that offer the maximal possible conclusions w.r.t. the selection criterion (namely *minsupp* and *minconf*), the retained premises for each FCT are then defined as follows.

Definition 10 In the context of our approach based on augmented Iceberg lattice, the set of valid premises w.r.t. the association rules that will be generated starting for a FCT T is defined as follows:

$$minG_T := \left\{ g \in allG_T \,\middle|\, \left(\nexists\, g_1 \in allG_T :\ g_1 \subset g \ \wedge\ \frac{Supp(T)}{Supp(g_1)} \geq minconf \right) \right.$$
$$\left. \wedge\ \left(\nexists\, s \in Cov^u(T) :\ \frac{Supp(s)}{Supp(g)} \geq minconf \right) \right\} \tag{13}$$

In other words, $minG_T$ only contains minimal, w.r.t. set inclusion, termsets leading to valid association rules based on the FCT T. On the other hand, the elements of $minG_T$ cannot be used as premises of valid association rules based on closed termsets subsuming T. Indeed, in this latter situation, rules based on T will not have the maximal possible conclusions given the associated premises.

Example 19 Consider the augmented Iceberg lattice depicted by Fig. 1 and Table 3. According to the previous example, we have $allG_{ACW} = \{A, C, W\}$. We will analyze the content of $minG_{ACW}$ with respect to different values of *minconf*.

- *Consider minconf = 1:* In this case, only A will be retained in $minG_{ACW}$. Indeed, A cannot be the premise of a larger conclusion than CW (*i.e.*, (ACW-A)). On the other hand, both other potential premises lead to rules which are not valid w.r.t. *minconf*. A unique rule based on ACW will then be retained, namely $A \stackrel{1}{\Rightarrow} CW$.
- *Consider minconf = 0.8:* In this case, both A and W will be retained in $minG_{ACW}$. Indeed, they cannot be used in rules with larger conclusions. The third potential premise, namely C, is not retained since it leads to a non valid rule ($Conf(C \Rightarrow AW) = 0.66 < 0.8$). Two rules based on ACW will then be retained, namely $A \stackrel{1}{\Rightarrow} CW$, and $W \stackrel{0.8}{\Rightarrow} AC$.
- *Consider minconf = 0.6:* In this case, only C will be retained in $minG_{ACW}$ since it leads to a valid rule, $C \stackrel{0.66}{\Rightarrow} AW$, while not being in the premise of a larger conclusion. Noteworthily, although A and W will lead to valid rules based on ACW, they are not retained since they are used in retained rules of larger conclusions (based on the FCT ACTW instead of ACW). These latter rules are: $A \stackrel{0.75}{\Rightarrow} CTW$, and $W \stackrel{0.6}{\Rightarrow} ACT$.
- *Consider minconf = 0.5:* In this case, the set $minG_{ACW}$ will be empty and there is no retained rule based on ACW. Indeed, although, the different potential premises lead to valid rules based on ACW, none of them is retained in $minG_{ACW}$. The reason is that they are used to give valid rules of larger conclusions based on the FCT ACTW, which are: $A \stackrel{0.75}{\Rightarrow} CTW$, $C \stackrel{0.5}{\Rightarrow} ATW$, and $W \stackrel{0.6}{\Rightarrow} ACT$.

The following propositions introduce *approximate and exact irredundant association rules* according to our generic basis.

Proposition 1 *Let T be a* FCT. *A valid irredundant approximate association rule mined starting from T is of the form $R: g \Rightarrow T$ with $g \in minG_T$ and $\Omega(g) \subset T$.*

Example 20 Consider the results obtained in Example 19 for *minconf* = 0.6. The rule $A \Rightarrow CTW$ having a confidence equal to 0.75 is a valid irredundant *approximate* rule. In this respect, $\Omega(A) = ACW \subset ACTW$.

Another way to express Proposition 1 is that there is not a FCT $s \in Cov^u(T)$ that can lead to a valid approximate association rule when g is taken as the premise of a rule based on s (*i.e.*, the rule $R : g \Rightarrow (s - g)$ is not valid for all $s \in Cov^u(T)$).

Proposition 2 *Let T be a* FCT. *An irredundant exact association rule mined starting from T is of the form $R: g \Rightarrow T$ with $g \in minG_T$ and $\Omega(g) = T$, i.e., $g \in G_T$.*

Example 21 Consider now the results obtained in Example 19 for $minconf = 0.8$. The rule A $\Rightarrow$ CW having a confidence equal to 1 is a valid irredundant *exact* rule. In this respect, $\Omega(\text{A}) = \text{ACW}$ (since A is a minimal generator of ACW).

According to both previous propositions, the set $min\mathcal{G}_T$ is partitioned into two parts: the first contains minimal generators having for closure a proper subset of T and, as a consequence, leading to valid approximate rules, while the second part gathers minimal generators having for closure T and, then, leading to exact rules. Note that an exact association rule is necessarily valid since it has a value of confidence equal to 1.

3.3.2 Definition of the minimal generic base $\mathcal{MGB}$

The proposed generic basis is defined as follows.

Definition 11 Given $\mathfrak{M} := (\mathcal{C}, \mathcal{T}, \mathcal{I})$, a textual context, T a frequent closed termset and its $min\mathcal{G}_T$, the minimal generic basis $\mathcal{MGB}$ is defined as follows:

$$\mathcal{MGB} := \{R : g \Rightarrow (T - g) | T \in \mathcal{FCT} \ \wedge \ g \in min\mathcal{G}_T\} \tag{14}$$

According to equation (14), Propositions 1 and 2, irredundant approximate rules (IARs) of the form $g_1 \Rightarrow (T_2 - g_1)$ link a minimal generator g_1 of a FCT T_1 and a second FCT T_2 such that $T_1 \subset T_2$ (*i.e.*, g_1 "implies" T_2, which is located higher in the augmented Iceberg lattice, with a confidence equal to $\frac{Supp(T_2)}{Supp(g_1)}$). On the other hand, the derived irredundant exact rules (IERs) have the following form: $g \Rightarrow (\Omega(g) - g)$, given that g does not appear as a premise of any other valid approximate rule of a conclusion larger than $(\Omega(g) - g)$. For example, for $minconf = 0.6$, the rule A $\Rightarrow$ CW is not retained among IERs of $\mathcal{MGB}$ since it is redundant w.r.t. the IAR: A $\Rightarrow$ CTW (*cf.* Example 19).

We introduce in what follows the Gen-MGB algorithm which allows the construction of the $\mathcal{MGB}$ generic basis.

3.3.3 Description of the Gen-MGB *algorithm*

In our approach, the augmented Iceberg lattice supports the irredundant association rules discovery between terms. The main advantage brought by this partially ordered structure is the efficiency. In fact, by using such a precedence order, irredundant exact and approximate association rules are directly derived, without additional confidence measure computations.

The pseudo-code of the Gen-MGB algorithm is given by Algorithm 1. It iterates on the set of frequent closed termsets $\mathcal{FCT}$ of the augmented Iceberg lattice $\mathcal{AL}$, starting from larger FCTs and sweeping downwardly w.r.t. set inclusion $\subseteq$.

The algorithm takes the augmented Iceberg lattice $\mathcal{AL}$ as input and gives as output the irredundant approximate and exact association rules (*i.e.*, IARs and IERs). With respect to Proposition 1 and considering a given node in the Iceberg lattice, we consider that IARs represent implications that involve the minimal generators of the sub-closed-termset, associated to the considered node, and a super-closed-termset. On the other hand, w.r.t. Proposition 2, IERs are implications extracted

using minimal generators and their respective closures, belonging to the same node in $\mathcal{AL}$.

The generation of irredundant association rules with GEN-MGB algorithm is performed in two steps, described as follows.

Step 1: Candidate Conclusions Generation

The goal of this step is to find, for a given FCT T_i, those FCTs which represent the candidate conclusions for association rules having g_i as premise, where $g_i \in \mathcal{G}_{T_i}$. The targeted FCTs are those which include the FCT T_i under consideration. Thus, an association rule R between the minimal generators g_i of T_i and a FCT T_j is valid if and only if:

$$Conf(R) = \frac{Supp(T_j)}{Supp(g_i)} = \frac{Supp(T_j)}{Supp(T_i)} \geq minconf \tag{15}$$

Equation (15) means that $Supp(T_j) \geq Supp(T_i) \times minconf$. As a first filter on the list of FCTs subsuming T_i, we use the value *threshold-Supp* which is equal to $(minconf \times Supp(T_i))$ to only retain those FCTs having a support greater than or equal to *threshold-Supp*. Thus, instead of computing for each FCT T_j of the input list the confidence given by the equation (15), the GEN-MGB algorithm only checks whether $Supp(T_j) \geq$ *threshold-Supp*.

The GET-CONCLUSION function adds the FCT $T_j \supset T_i$ (line 7) to the set *limit-closures* if T_j is a maximal FCT, w.r.t. set inclusion, amongst those FCTs retained after the use of the aforementioned filter. The set *limit-closures* gathers FCTs based on them will be derived valid association rules having as premise $g_i \in \mathcal{G}_{T_i}$, while considering the conditions specified in Definition 10.

Algorithm 1: The GEN-MGB algorithm

```
1:  Algorithm GEN-MGB(In: AL: the augmented Iceberg lattice, Out: The minimal
    generic basis MGB)
2:  for all FCT T_i ∈ AL do
3:     threshold-Supp ← minconf × Supp(T_i)
4:     limit-closures ← ∅
5:     for all FCT T_j such that T_i ⊂ T_j do
6:        if Supp(T_j) ≥ threshold-Supp then
7:           GET-CONCLUSION(T_j, limit-closures)
8:     end for
9:     if limit-closures ≠ ∅ then
10:       GENERATE-IARS(limit-closures, T_i, IARs)
11:    else
12:       GENERATE-IERS(T_i, IERs)
13: end for
14: return (MGB = {IARs ∪ IERs})
```

Step 2: Irredundant Association Rules Generation

During this step, for each FCT T_i, two cases are distinguished according to the content of its *limit-closures* list:

1. If the *limit-closures* list is not empty (line 9), the algorithm generates IARs. To only retain non-redundant rules, the GEN-MGB algorithm manages a *prohibited*

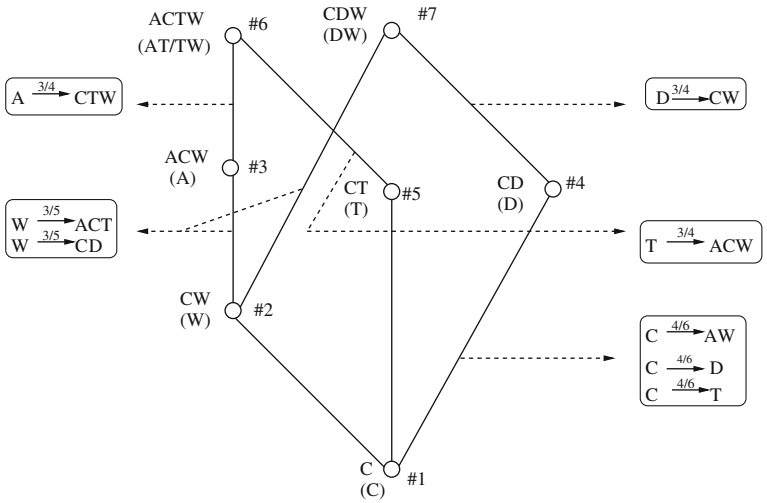

Fig. 2 Irredundant association rules associated to the textual context $\mathfrak{M}$

list for each FCT containing already retained premises of rules based on this FCT. In this regards, before generating new rules, Gen-MGB checks whether a proper subset of the termset $g_i \in \mathcal{G}_{T_i}$ is already in the *prohibited list* of T_j ($T_j \in$ *limit-closures*). If it is the case, this means that the rule based on T_j having g_i for premise is redundant w.r.t. another one already extracted. Otherwise, the approximate rule $g_i \Rightarrow (T_j - g_i)$ is mined. The termset g_i is added to the *prohibited list* of the FCT T_j.

2. In the case where the Get-Conclusion function returns an empty list, *i.e.*, the *limit-closures* set is empty (line 11), the algorithm generates the IERs, associated to the FCT T_i. Each derived exact rule takes the form $g_i \Rightarrow T_i - g_i$.

Example 22 Consider the augmented Iceberg lattice depicted in Fig. 1 for *minconf* = 0.6. Let us recall that the set value of *minsupp* is equal 3. All irredundant approximate association rules are depicted in Fig. 2. In this case, none irredundant exact rule is mine since all of them are redundant w.r.t. irredundant approximate rules belonging to $\mathcal{MGB}$. For example, starting from the node having CDW for frequent closed termset, the exact rule DW $\overset{1}{\Rightarrow}$ C is not generated since it is considered as redundant w.r.t. the approximate association rule D $\overset{0.75}{\Rightarrow}$ CW, according to Definition 8.

3.3.4 Redundant association rule derivation

Our generic basis allows deriving all valid redundant association rules thanks to two derivation axioms of the inference mechanism, defined and proven to be sound and complete in BenYahia et al. (2009), as follows:

1. *Augmentation:* if $T_1 \Rightarrow T_2 \in \mathcal{MGB}$ and $T_3 \subset T_2$ then $T_1 \cup T_3 \Rightarrow T_2 - T_3$ is a valid association rule.

2. *Conditional decomposition:* if $T_1 \Rightarrow T_2 \in \mathcal{MGB}$ and $T_3 \subset T_2$, then $T_1 \Rightarrow T_3$ is a valid association rule.

Note, however, that $\mathcal{MGB}$ is not informative. For example, the support and confidence of the rule AT $\Rightarrow$ CW would not be derived from A $\Rightarrow$ CTW. This results from the fact that $\mathcal{MGB}$ rules do not necessarily allow to locate the corresponding frequent closed termset of a given termset, *i.e.*, its closure. This is the case for the termset AT what makes its support not exactly known.

Interestingly enough, the fact that $\mathcal{MGB}$ is not informative will not affect the quality of our expansion process. Indeed, the confidence measure is used in our approach as a filter to eliminate non valid rules. Once irredundant rules retained in $\mathcal{MGB}$, their respective confidence values are no more taken into account in the proposed expansion process. Our goal is in fact to expand an original query with as few rules as possible, *i.e.*, those of $\mathcal{MGB}$ fulfilling some constraints imposed by the selection criteria. This is carried out through the maximal possible set of terms characterizing each applied rule of $\mathcal{MGB}$. It is important to note that in the experiments we carry out, we analyze the effect of the *minconf* value on the results of the query expansion process. This will prove that rule of higher confidence values (like exact ones) do not necessarily lead to better results than lower confidence rules. These latter rules have the advantage of linking terms that do not always co-occur in the same documents, *i.e.*, inter-document relationships. This is not possible using very strong rules mainly conveying intra-document relationships.

3.4 Theoretical comparison of generic bases of association rules

In this subsection, we will further compare our minimal generic basis $\mathcal{MGB}$, by applying the Gen-MGB algorithm on the running example given in Table 3, with

Table 4 Rules discovered from the running example context for *minsupp* = 3 and *minconf* = 0.6: *(Left)* Rules of $\mathcal{MGB}$, *(Center)* Rules discovered by the Bastide et al.'s approach, and *(Right)* Rules discovered by the Zaki's approach

$\mathcal{MGB}$	$\mathcal{GBA}$	$\mathcal{GBE}$	$\mathcal{NRR}$	
C $\overset{4/6}{\Rightarrow}$ AW	C $\overset{5/6}{\Rightarrow}$ W	D $\overset{1.0}{\Rightarrow}$ C	C $\overset{4/6}{\Rightarrow}$ D	TW $\overset{1.0}{\Rightarrow}$ A
C $\overset{4/6}{\Rightarrow}$ T	C $\overset{4/6}{\Rightarrow}$ AW	T $\overset{1.0}{\Rightarrow}$ C	C $\overset{4/6}{\Rightarrow}$ T	A $\overset{1.0}{\Rightarrow}$ W
C $\overset{4/6}{\Rightarrow}$ D	C $\overset{4/6}{\Rightarrow}$ T	W $\overset{1.0}{\Rightarrow}$ C	C $\overset{5/6}{\Rightarrow}$ W	A $\overset{1.0}{\Rightarrow}$ C
W $\overset{3/5}{\Rightarrow}$ ACT	C $\overset{4/6}{\Rightarrow}$ D	A $\overset{1.0}{\Rightarrow}$ CW	C $\overset{4/6}{\Rightarrow}$ A	W $\overset{1.0}{\Rightarrow}$ C
W $\overset{3/5}{\Rightarrow}$ CD	W $\overset{4/5}{\Rightarrow}$ AC	DW $\overset{1.0}{\Rightarrow}$ C	W $\overset{4/5}{\Rightarrow}$ A	T $\overset{1.0}{\Rightarrow}$ C
T $\overset{3/4}{\Rightarrow}$ ACW	W $\overset{3/5}{\Rightarrow}$ ACT	AT $\overset{1.0}{\Rightarrow}$ CW	W $\overset{3/5}{\Rightarrow}$ D	D $\overset{1.0}{\Rightarrow}$ C
D $\overset{3/4}{\Rightarrow}$ CW	W $\overset{3/5}{\Rightarrow}$ CD	TW $\overset{1.0}{\Rightarrow}$ AC	W $\overset{3/5}{\Rightarrow}$ T	
A $\overset{3/4}{\Rightarrow}$ CTW	T $\overset{3/4}{\Rightarrow}$ ACW		A $\overset{3/4}{\Rightarrow}$ T	
	D $\overset{3/4}{\Rightarrow}$ CW		D $\overset{3/4}{\Rightarrow}$ W	
	A $\overset{3/4}{\Rightarrow}$ CTW		T $\overset{3/4}{\Rightarrow}$ A	
			T $\overset{3/4}{\Rightarrow}$ W	

those proposed respectively by Bastide et al. (2000) and Zaki (2004). The obtained results are given by Table 4 and allow us to draw the following observations:

– In the general case, it is important to note that $\mathcal{MGB}$ rules are necessarily in Bastide et al. generic basis. Indeed, if a rule R belonging to $\mathcal{MGB}$ is exact, then $R \in \mathcal{GBE}$, while if it is approximate, then $R \in \mathcal{GBA}$. This can clearly be explained by the fact that the rules of $\mathcal{MGB}$ as well as the couple ($\mathcal{GBE}$, $\mathcal{GBA}$) are of minimal premise and of maximal conclusion. The advantage of $\mathcal{MGB}$ is that, for a given premise, it only looks for the maximal possible conclusions (*i.e.* those incomparable w.r.t. set inclusion). For example, considering W as a premise. $\mathcal{MGB}$ only contains rules W $\stackrel{3/5}{\Rightarrow}$ ACT and W $\stackrel{3/5}{\Rightarrow}$ CD (*cf.* Table 4). The maximal possible conclusions are in fact ACT and CD which are incomparable w.r.t. set inclusion. Each conclusion will then add at least a new term to the expanded query that can not be added using another rule.
On the other hand, ($\mathcal{GBE}$, $\mathcal{GBA}$) contains all generic rules of maximal conclusion having a given premise. Thus, for the premise W, not only it contains those rules contained in $\mathcal{MGB}$ but also the rules W $\stackrel{4/5}{\Rightarrow}$ AC and W$\stackrel{1.0}{\Rightarrow}$ C. In this situation, four conclusions are found to give valid generic rules of premise W. However, since C $\subset$ CD and AC $\subset$ ACT, both conclusions C and AC are not retained in $\mathcal{MGB}$ because they are not maximal possible ones.
We then have in the general case: $\mathcal{MGB} \subseteq \mathcal{GBE} \cup \mathcal{GBA} \subseteq \mathcal{VAR}$. While the three sets $\mathcal{MGB}$, ($\mathcal{GBE} \cup \mathcal{GBA}$), and $\mathcal{VAR}$ allow expanding a given query by the same set of terms, the main advantage of $\mathcal{MGB}$ is a very reduced set of rules thus avoiding the use of a large amount of redundant rules what is important for improving running time of the query process as well as optimizing memory requirements for storing retained rules.
The definition of *Bastide et al.* (*cf.* Section 3.1) thus suffers from high rigidity. Actually, having exactly the same support and the same confidence as elimination criteria fails to exclude rules of identical premises having comparable conclusions w.r.t. set inclusion. This leads to the generation of a large number of rules, many of which convey the same query expansion possibilities in IR field. For example, from Table 4 (Center), this approach keeps both the rules C $\stackrel{5/6}{\Rightarrow}$ W and C $\stackrel{4/6}{\Rightarrow}$ AW. While our approach will only keep the last one. The motivation behind retaining the last rule is that it brings more additional terms to an expanded query, in comparison to the first one.
For example, in query expansion process, it is useless to keep the rule *mining* $\Rightarrow$ *termsets*, provided that another rule, say *mining* $\Rightarrow$ *document* $\wedge$ *term* $\wedge$ *termset* $\wedge$ *frequent* is already available. The latter is more interesting, since by associating more terms to the term *mining*, it has more chances of increasing the results quality of an information retrieval system including this term.
– On the other hand, *Zaki* introduced an axiomatic-based approach for deriving a generic basis for the entire rule set (*cf.* Section 3.2), without any claim of minimality, *i.e.*, compactness of this basis. Note that the author considered only rules among neighbor closed termsets in the Iceberg lattice, and claimed that the remaining rules can be inferred by applying the transitivity axiom to that basis and the information about the order in this ordered structure (Zaki 2004).

An axiomatic system is used in order to minimize the size of both the premise and conclusion parts of the rules. However, this will lead to the elimination of rules having larger and therefore more interesting conclusion parts in favor of rules having minimal conclusions. Of course, the same information conveyed in a larger rule can be extracted from a set of smaller ones by applying the transitivity axiom (Luxenburger 1991). For example, from the output of the Zaki's method, illustrated in Table 4 (Right), we conclude that even though the rule $C \overset{4/6}{\Rightarrow} AW$ is not part of the output of Zaki's method, it can still be discovered by composing $C \overset{5/6}{\Rightarrow} W$ and $W \overset{4/6}{\Rightarrow} A$.

For a structural view point, in the general case, the rules of $\mathcal{MGB}$ can not be compared to those of $\mathcal{NRR}$. Indeed, although both generic bases offer rules of minimal premises, the former conveys rules of maximal conclusions, while the latter those of minimal conclusions. However, as shown in Table 4, a unique rule of $\mathcal{MGB}$ often covers several ones of $\mathcal{NRR}$ as it is the case of $A \Rightarrow CTW$ w.r.t. the rules $A \Rightarrow C$, $A \Rightarrow T$, and $A \Rightarrow W$ of $\mathcal{NRR}$.

Similarly, the rule *mining* $\Rightarrow$ *termset* $\wedge$ *frequent* $\wedge$ *transaction* $\wedge$ *attribute* will be certainly discarded in profit of minimal rules such as *mining* $\Rightarrow$ *termset* and *termset* $\Rightarrow$ *frequent*, etc. The discarded rule can be obtained by several chaining steps, corresponding to transitivity axiom of Luxenburger (1991) and that of the augmentation of Armstrong (1974). However, the application of Zaki's approach to query expansion will require an additional rule expansion as preprocessing step. Consequently, the cost of every single query expansion increases and may make the whole process too inefficient in case of a long run.

3.5 Experimental comparison of generic bases of association rules

To evaluate the reduction rate offered by $\mathcal{MGB}$ of the whole set of valid association rules, we carried out different experiments in four steps described as follows:

1. In order to extract the set of the frequent closed termsets, *i.e.*, $\mathcal{FCT}$ associated to their respective minimal generators, we adapted the GC-Growth algorithm (Haiquan et al. 2005) to our textual context.
2. Irredundant approximate and exact association rules between terms are then generated from the set $\mathcal{FCT}$ of frequent closed termsets, *i.e.*, the $\mathcal{MGB}$ generic basis, using the Gen-MGB algorithm described below.
3. We generated the whole set of valid association rules, *i.e.*, that also including redundant ones, further denoted by $\mathcal{VAR}$, using the implementation of the Apriori algorithm (Agrawal and Skirant 1994) of Bart Goethals.[7]
4. The reduction rate offered by $\mathcal{MGB}$ w.r.t. the whole valid association rule set $\mathcal{VAR}$, is then computed over the different document collections, using the following formula:

$$Reduction\_rate = \frac{|\mathcal{VAR}| - |\mathcal{MGB}|}{|\mathcal{VAR}|} \tag{16}$$

Equation (16) represents the compression rate of the whole set of valid association rules $\mathcal{VAR}$.

[7] Available at: http://www.adrem.ua.ac.be/~goethals/software/.

Table 5 Characteristics of the used document collections

Campaign	Collection	Size (Mb)	# documents	# distinct terms	# queries
Amaryllis II	OFIL	≈ 33	11,016	119,434	26
	INIST	≈ 68	163,307	174,659	30
CLEF 2003	Le Monde 94	≈ 158	44,013	106,558	50
	ATS 94	≈ 86	43,178	55,526	50
	Le Monde 94 & ATS 94	≈ 244	87,191	113,422	50

3.5.1 Description of the used collections

Experiments are conducted on five document collections, namely:

- The OFIL and INIST document collections of the second Amaryllis campaign.
- The Le Monde 94 and ATS 94 document collections of the CLEF 2003 campaign (Collection 2001). We also used a fifth collection composed of both Le Monde 94 documents and ATS 94 documents.

A set of queries is associated to each collection and, for each query, a set of relevant documents is assigned. Table 5 gives further details about the used collections.

We notice that the OFIL, Le Monde 94 and ATS 94 document collections are composed of articles from national French newspapers, while the INIST document collection contains abstracts of scientific papers caught from the *PASCAL* (during four years) and *FRANCIS* (during one year) collections. From the view point of collection characteristics, the first three collections have fewer documents than the last one. However, the documents in the INIST collection are shorter than those of the other collections.

3.5.2 Document collections preprocessing

In order to extract the most representative terms, a linguistic preprocessing is performed on the document collections by using the French morpho-syntactic tagger Cordial.[8] In this application, we focus only on terms related to two grammatical categories: *the common nouns* and *the proper nouns*. A stoplist is used to discard functional French terms that are very common, *e.g.*, *le*, *la*, *donc*, etc.

The context document-term $\mathfrak{M}$ is then built by retaining only terms corresponding to the selected grammatical categories. The association rules are then generated from the augmented Iceberg lattice using the Gen-MGB algorithm. The minimal threshold of confidence is set to 50% and we varied the minimum and maximum threshold of the support, *i.e.*, *minsupp* and *maxsupp*,[9] w.r.t. the document collection size and to term distributions. While considering the *Zipf* distribution of every collection, the maximum threshold of the support values is experimentally set in order to spread trivial terms which occur in the most of the documents, and are then related to too many terms. On the other hand, the minimal threshold allows eliminating marginal terms which occur in few documents, and are then not statistically important when occurring in a rule.

[8]Distributed by the *Synapse Development* Corporation.

[9]*maxsupp* means that the termset must occur at most below than this user-defined threshold.

3.5.3 Reduction results

Table 6 summarizes the number of mined association rules between terms as well as the reduction rate for the different support intervals, *i.e.*, [*minsupp*, *maxsupp*] intervals, respectively to the considered collections.

We observe that for the OFIL collection, an important number of associations between terms were discovered for a support interval between 5 and 1,000 documents. This fact is not in contradiction with the OFIL terms distribution, where the support of terms is important between 25 and 200 documents. Note also that for the INIST collection, an important number of associations between terms have been discovered for a support interval between 25 and 500 documents, which is justified by the high term support in INIST collection in this support interval.

On the other hand, the Le Monde 94 document collection behaves as a "worst case" context, w.r.t. the Galois closure operator, where each frequent closed termset (FCT) is exactly equal to its minimal generator. This arises for almost all tested *minsupp* values, and even for very low ones. For example, for *minsupp* = 150, there is as many FCTs as many minimal generators, equal to 310, 181. In addition, each frequent termset is in the general case simultaneously closed and minimal generator, except four termsets (on 310,185). As a consequence and contrary to both previous document collections, a very few number of *exact* association rule with a *non-empty* conclusion is generated starting from this collection since each frequent minimal generator g is itself a FCT. A rule of the form $g \Rightarrow \emptyset$ is in fact commonly considered as non-informative. From the expansion process point of view, such a rule is not of any added-value since the presence of the terms in the premise, those of g, does not imply that of other terms. Interestingly enough, the search space associated to this collection tends to be an antimatroid closure space (Pfaltz and Taylor 2002), which corresponds to the case where each subset of objects of the extraction context (the documents in our case) shares a common pattern—a closed termset—which is itself a minimal generator.

Table 6 Reduction results on OFIL, INIST, Le Monde 94 and ATS 94 document collections (All valid rules versus $\mathcal{MGB}$)

Support interval	Size($\mathcal{VAR}$)	Size($\mathcal{MGB}$)	Reduction rate (in %)
OFIL			
[5, 50] documents	235,806	5,761	97.56
[50, 1,000] documents	291,062	85,878	70.49
[1,000, 5,000] documents	374	257	31.28
INIST			
[3, 30] documents	5,154	3,062	40.59
[30, 250] documents	472	273	42.16
[250, 16,000] documents	11,012	8,949	18.73
Le Monde 94			
[150, 1,500] documents	1,965,766	716,842	63.53
[200, 2,000] documents	709,904	300,493	57.67
[300, 3,000] documents	171,846	89,072	48.17
ATS 94			
[150, 1,500] documents	113,220	42,393	62.56
[200, 2,000] documents	44,574	22,297	49.98
[300, 3,000] documents	15,624	8,816	43.57

Table 7 Characteristics of the tested benchmark datasets

Dataset	Number of transactions	Number of items	Average size of transactions	Maximal size of transactions
Connect	67,557	129	43.00	43.00
Chess	3,196	75	37.00	37.00
Mushroom	8,124	119	23.00	23.00
Retail	88,162	16,470	10.31	77.00

The ATS 94 collection behaves in a similar manner than the previous collection. Indeed, the number of minimal generators is always almost equal to that of FCTs. For example, for *minsupp* = 100, we have 60,709 FCTs while the number of their associated minimal generators is equal to 60,949. The difference is then equal to 240, which constitutes less than 0.40% of the whole number of minimal generators. Note also that for this very low *minsupp* value, the number of frequent termsets is equal to 62,847. Thus, the number of frequent termsets which are not closed does not exceed 3.52% of the whole number of frequent termsets. It results from such characteristics of this collection a very reduced number of exact association rules appear with a non-empty conclusion.

3.5.4 Compression rates of $\mathcal{MGB}$ versus other generic bases

For the sake of comparing the compression rates offered by $\mathcal{MGB}$ to those of the other generic bases, we now present results of experiments we carried out on benchmark datasets frequently used for assessing performances of approaches dedicated to association rule mining.[10] The characteristics of the tested benchmark datasets are given in Table 7. The choice of these datasets is argued by the fact that the extraction tools of the generic bases of the literature are not designed for large extraction contexts, especially w.r.t. to the number of distinct terms, as it is the case for the document collections we used in our work. In particular, these tools are not dedicated to datasets associated to the text mining field. This explains the optimizations we introduced in the design and the implementation of the proposed Gen-MGB algorithm to make it a dedicated tool for mining generic rules of $\mathcal{MGB}$ from texts.

The obtained results are depicted in Table 8 and allow pointing out the interesting compression rates of the proposed $\mathcal{MGB}$. In this respect, $RR_{\mathcal{B}}$ denotes the reduction rate, given in percentage, offered by $\mathcal{MGB}$ versus another set of rules (*i.e.*, either ($\mathcal{GBE} \cup \mathcal{GBA}$), $\mathcal{NRR}$, or $\mathcal{VAR}$). The used formula is the same as equation (16) by substituting $\mathcal{VAR}$ by the corresponding set of rules.

As discussed in Section 3.4, we always have: $Size(\mathcal{MGB}) \leq Size(\mathcal{GBE} \cup \mathcal{GBA}) \leq Size(\mathcal{VAR})$. Thus, $\mathcal{MGB}$ allow removing a large amount of redundant association rules.

In addition, although the generic basis $\mathcal{NRR}$ is smaller than ($\mathcal{GBE} \cup \mathcal{GBA}$), the proposed generic basis in this work is even more reduced than $\mathcal{NRR}$ and this occurs for all tested benchmark datasets. This can be explained by the fact that each rule of $\mathcal{MGB}$ has a maximal conclusion which is not the case for a rule of $\mathcal{NRR}$ having a minimal conclusion. Thus, a rule of $\mathcal{MGB}$ often covers one or more rules of $\mathcal{NRR}$

[10]Test datasets are available at: http://fimi.cs.helsinki.fi/data.

Table 8 Reduction results on benchmark datasets ($\mathcal{MGB}$ versus ($\mathcal{GBE}$, $\mathcal{GBA}$) and $\mathcal{NRR}$)

minconf (in %)	Size ($\mathcal{MGB}$)	Size ($\mathcal{GBE} \cup \mathcal{GBA}$)	Size ($\mathcal{NRR}$)	Size ($\mathcal{VAR}$)	$RR_{(\mathcal{GBE}, \mathcal{GBA})}$	$RR_{\mathcal{NRR}}$	$RR_{\mathcal{VAR}}$
CONNECT (*minsupp* = 64,179)							
95.00	635	25,336	3,054	77,816	97.49	79.20	99.18
96.00	852	18,470	3,054	73,869	95.39	72.10	98.85
97.00	1,403	18,470	3,051	60,101	92.40	54.02	97.67
98.00	1,033	11,717	2,804	41,138	91.18	63.16	97.49
99.00	1,386	5,250	2,473	19,967	73.60	43.95	93.06
CHESS (*minsupp* = 2,780)							
87.00	440	31,538	4,873	42,740	98.60	90.97	98.97
89.00	519	29,704	4,873	40,451	98.25	89.35	98.72
91.00	627	26,147	4,799	36,098	97.60	86.93	98.26
93.00	793	21,350	4,734	29,866	96.29	83.25	97.34
95.00	671	14,373	4,260	20,312	95.33	84.25	96.70
MUSHROOM (*minsupp* = 2,437)							
30.00	332	7,623	1,829	94,894	95.64	81.89	99.65
50.00	366	5,761	1,732	79,437	93.65	78.87	99.54
70.00	364	2,159	1,501	58,010	83.14	75.75	99.37
90.00	498	2,159	933	24,408	76.93	46.62	97.96
RETAIL (*minsupp* = 440)							
00.50	435	1,382	861	1,382	68.52	49.48	68.52
01.00	402	1,334	838	1,334	69.86	52.03	69.86
10.00	232	770	553	770	69.87	58.05	69.87
50.00	304	438	402	438	30.59	24.39	30.59

which results in a size of $\mathcal{MGB}$ which is lower than that of $\mathcal{NRR}$ for the different tested datasets.

These experiments prove that $\mathcal{MGB}$ constitutes a very reduced set of rules that allows covering all remaining valid redundant association rules.

4 Automatic query expansion based on association rules

In this section, we firstly discuss the main related works to our approach of automatic query expansion based on association rules. Then, we present in details the key notions of our expansion process.

4.1 Related works

Query expansion aims to improving a user's search by adding new query terms to an existing query either by the user, commonly called interactive query expansion (IQE) (Joho et al. 2004), or by the retrieval system, commonly called automatic query expansion (AQE). Recently, Ruthven showed that automatic query expansion (AQE) can better reach this aim than the interactive one because human searchers are less likely than systems to make good expansion decisions (Ruthven 2003).

Different methods dedicated to query expansion have been proposed in the literature such as those based on syntactic context (Bodner and Song 1996; Grefenstette 1992), user relevance feedback (Ruthven and Lalmas 2003; Schenkel and Theobald

2005), pseudo relevance feedback (Buckley et al. 1994; Mitra et al. 1998), and terms co-occurrences (Lin et al. 2008; Rungsawang et al. 1999; Sun et al. 2006).

With respect to the syntactic context, Grefensette combines in Grefenstette (1992) static co-occurrences and the head-modifier relation between terms to extract related words where the window size is a sentence.

In a user relevance feedback context, related terms come from user identified relevant documents or queries. One of these techniques is the use of web servers or Web search engines query logs file for mining related queries (Shi and Yang 2007; Yurekli et al. 2009). The added terms are related queries (or subset of related queries) based in the query logs of previously submitted queries by human users identified using association rules. The expansion can be done either automatically or users can use the suggested related queries to modify their original ones. Fonseca et al. (2005) segmented query sessions in search engine query logs into subsessions and then used association rules to extract related queries from those subsessions. They calculated the relatedness between all queries using the association rule mining model and then built a query relation graph. The query relation graph was used for identifying terms related to a given user input query.

On the other hand, the pseudo relevance feedback expanded terms which come from the top k retrieved documents assumed to be relevant without any intervention from the user. The authors in Buckley et al. (1994) and Mitra et al. (1998) proposed approaches for expanding search engine queries. The related terms are extracted from the top documents that are returned in response to the original query using statistical heuristics, and the query is expanded using these extracted terms. The results of this approach on large collections are sometimes even negative since the assumed relevant documents retrieved by an information retrieval system are unfortunately not all relevant (Buckley et al. 1994). Another limitation of this technique is that it is using a local query expansion technique based on a set of documents retrieved for the query. As a consequence, they are more focused on the given query than global analysis. Indeed, in Xu and Croft (1996), the authors showed that using global analysis techniques produces results that are both more effective and more predictable than simple local feedback.

The reuse of an existent external knowledge sources, such as Wordnet (Voorhees 1993) and more recently Wikipedia (Hu et al. 2009), has produced few successful experiments. It is not clear how to adapt the external knowledge source structure to IR and query term disambiguation.

Moreover, the authors in Croft and Yufeng (1994) exploited lexical co-occurrence to build a local document collection thesaurus where the window size is a paragraph (containing from 3 to 10 sentences).

Association rules techniques extract relationships based on their co-occurrences where the window size used is a document. The authors of Tangpong and Rungsawang (2000) performed a small improvement when using the Apriori algorithm (Agrawal and Skirant 1994) with a high confidence threshold (more than 50%) that generated a small amount association rules. Using a lower confidence threshold (10%), authors performed better results (Rungsawang et al. 1999). The same approach is proposed by Haddad et al. (2000) performing improvement when using the Apriori algorithm to extract association rules. The best improvements were performed with low confidence values. The main limitation of this approach consists in the huge number of generated association rules while a large part of them are

redundant in the sense that several rules convey the same information. The removal of redundancy within mined rules is then a key step for improving the quality of the expansion as performed in the approach we propose in this work.

In comparison to the related work, our approach is based on a co-occurrence technique that extracts relations between terms based on a global analysis of a document collection. This is carried out through association rules which convey statistical relations between terms used in an automatic query expansion process as detailed in the following.

4.2 Automatic query expansion based on generic basis of association rules

The purpose of this subsection is to disclose how that can be achieved when a query is expanded using association rules. We will compare the effectiveness of the expanded queries with that of original queries.

Our approach is based on terms co-occurrences. In this regard:

- Words can co-occur when they are adjacent or when they are separated by a number of intervening words. The distance between words considered to co-occur is called the window size. In almost all existing works based on words co-occurrence, the considered window size is the sentence where words co-occur or the number of intervening words. In our context, the window size is the document.
- The filtering is carried out w.r.t. the minimum support threshold *minsupp* and the selected grammatical categories.

Our automatic Query Expansion approach can thus be regarded as a more elaborated co-occurrence based technique. Indeed, an association rule conveys a global relation between terms which does not depend on a given document but implies a set of documents characterizing a group of linked terms. In this respect, more than a local co-occurrence based technique only allowing to guess relations between terms of a given document—intra-document relationship—an association rule offers information about inter-document relationships of terms and intra-document relationships. Association rules thus offer finer relations between the terms than simpler approaches. It is also interesting to note that our query expansion approach does not require a priori knowledge or a complicated linguistic process. It is based on an automatic process without any external or human intervention neither external knowledge resources (thesaurus, ontology, etc.).

4.2.1 Steps of the query process

The process of automatic query expansion is handled on three steps, namely:

1. *Baseline run:* Finding the best performances of the OQ. These latter are measured by:
   - Precision of the original query (OQ) set at eleven representative recall points (P@11). This precision measure takes all retrieved documents into account.
   - P@5, P@10, P@15, and P@30 documents are precision measures evaluated at a given cut-off rank, considering only the top most results returned by the system. For example, P@5 is precision at 5 documents.
   - Mean average precision (MAP).

2. *Automatic query expansion:* Expanding each original query of the collection by all terms that appear in the conclusions of the association rules whose associated premises are composed by terms of the original query. All query fields (title, description and narrative fields) are used during the automatic query expansion process.
3. *Second run:* a second run is launched with the expanded queries (EQ) and an evaluation is performed using the same measures as for the OQ. The obtained results are then compared to the baseline run. This allows assessing the effect of query expansion. Since lots of valuable information is considered, a significant improvement is expected. Note that the improvement value is computed as follows:

$$Improvement = \frac{(Result\ of\ the\ second\ run) - (Result\ of\ the\ baseline\ run)}{(Result\ of\ the\ baseline\ run)} \tag{17}$$

Hence, the proposed automatic query expansion process based on association rules between terms consists of expanding each query by all terms that appear in the conclusions of the irredundant association rules whose premise is contained by the original query. Each term of the query is handled individually.

Given an original query $OQ := \{t_1, t_2, \ldots, t_n\}$, the basic idea of our query expansion process for obtaining the associated expanded query EQ is defined as follows:

$$\forall R : T_1 \Rightarrow T_2, \text{an irredundant rule} \in \mathcal{MGB} : \ if\ T_1 \subseteq \text{OQ},\ then\ \text{EQ} := \text{OQ} \cup T_2. \tag{18}$$

Equation (18) means that if the premise of R is contained in the OQ, the terms of the conclusion will be added to the EQ. For a given OQ, the expansion process is then iteratively carried out for all rules of $\mathcal{MGB}$.

4.2.2 Evaluation framework

In our experiments, we used the Lemur toolkit.[11] To evaluate our approach for automatic query expansion, we used three weighting schemes, namely:

- $tf \times idf$ weighting[12] which is a family of vector based information retrieval schemes (Salton and Buckely 1988).
- BM25tf weighting scheme, one variant of the $tf \times idf$ model based on a probabilistic retrieval model.
- Okapi BM25 which is a probabilistic ranking model used by information retrieval systems to rank matching documents according to their relevance to a given search query (Jones et al. 2000).

[11] Freely available at: http://sourceforge.net/projects/lemur/, while the information about this toolkit is available at: http://www.lemurproject.org/.

[12] idf is the acronym of inverted document frequency.

5 Experimental results of automatic query expansion

We now discuss the experimental results based on irredundant association rules. The tested document collections are those described in Table 5, using the LEMUR toolkit.

5.1 Results of the automatic query expansion based on generic association rules

We used different sets of association rules generated by varying the minimum and the maximum support thresholds. Table 9 shows the best results for the different collections. In this table, "IARs" denotes valid irredundant approximate rules, while "IERs" denoted valid irredundant exact rules.

Table 9 highlights the retrieval quality difference between the original queries (Baseline Run) and the expanded ones using irredundant association rules of the $\mathcal{MGB}$ generic basis. These results are expressed in terms of the average precision at the 11 recall points for the different tested collections. For the three weighting schemes, our method yields an improvement of the average precision at 11 recall points where the best performances are found using the OKAPI BM25. The different

Table 9 Improvement of the average precision at 11 points of recall (P@11) for the different used collections in query expansion using $\mathcal{MGB}$ association rules. The minimum and maximum support thresholds are depicted in the following form: *$\mathcal{MGB}$-minsupp-maxsupp*

Evaluation	$tf \times idf$		BM25*tf*		OKAPI BM25	
Experiments	OFIL collection (AMARYLLIS II campaign)					
Baseline	24.83%		31.71%		31.44%	
$\mathcal{MGB}$-5-50		Improvement		Improvement		Improvement
Using IARs and IERs	32.06%	+29.11%	36.82%	+16.11%	37.61%	+19.62%
Using only IARs	29.59%	+19.17%	35.42%	+11.69%	35.73%	+13.64%
Using only IERs	28.12%	+13.25%	33.78%	+06.52%	34.09%	+08.42%
Experiments	INIST collection (AMARYLLIS II campaign)					
Baseline	15.47%		15.25%		15.48%	
$\mathcal{MGB}$-3-30		Improvement		Improvement		Improvement
Using IARs and IERs	17.34%	+12.21%	18.10%	+18.68%	18.91%	+22.15%
Using only IARs	17.32%	+12.92%	18.10%	+18.68%	18.90%	+22.09%
Using only IERs	17.36%	+12.21%	18.13%	+18.88%	18.91%	+22.15%
Experiments	LE MONDE 94 collection (CLEF 2003 campaign)					
Baseline	41.01%		42.56%		43.54%	
$\mathcal{MGB}$-300-3000		Improvement		Improvement		Improvement
Using IARs and IERs	42.59%	+03.85%	43.73%	+02.74%	45.04%	+03.44%
Experiments	ATS 94 collection (CLEF 2003 campaign)					
Baseline	53.03%		55.89%		56.56%	
$\mathcal{MGB}$-300-3000		Improvement		Improvement		Improvement
Using IARs and IERs	53.48%	+00.84%	55.92%	+00.05%	57.01%	+00.79%
Experiments	LE MONDE 94 & ATS 94 collections (CLEF 2003 campaign)					
Baseline	44.49%		47.51%		48.48%	
$\mathcal{MGB}$-300-3000		Improvement		Improvement		Improvement
Using IARs and IERs	45.45%	+02.15%	47.99%	+01.01%	49.24%	+01.56%

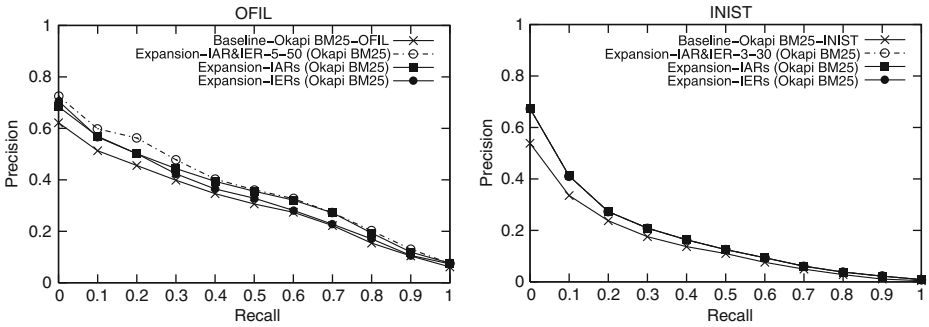

Fig. 3 Eleven points precision-recall curves for the tested document collections OFIL and INIST under the Okapi BM25 model

experiments carried out using $\mathcal{MGB}$ rules lead to an improvement w.r.t. those obtained using the baseline run. It is important to note that even for a large collection – the fifth one composed by the documents of Le Monde 94 and ATS 94 – we obtained interesting improvements.

Figure 3 sheds light on these important points. The use of all $\mathcal{MGB}$ rules often offers better results than the use of exact ones only (*cf.* the curves associated to the OFIL collection). This can be explained by the fact that approximate rules link terms that do not appear simultaneously in all cases, which can not be conveyed by exact rules. These latter only relate terms that always co-occur together. Note, however, that for the INIST collection, the different sets of irredundant rules used in

Table 10 Exact precision at 5, 10, 15 and 30 documents under the Okapi BM25 model

Evaluation	P@5 (%)	P@10 (%)	P@15 (%)	P@30 (%)
OFIL collection (Amaryllis II campaign)				
Baseline	40.77	37.69	33.33	27.31
$\mathcal{MGB}$-5-50	50.02	46.92	41.79	30.26
Improvement	+22.68	+24.48	+25.38	+10.80
INIST collection (Amaryllis II campaign)				
Baseline	32.67	30.00	28.22	21.44
$\mathcal{MGB}$-3-30	40.67	34.33	30.67	24.67
Improvement	+24.48	+14.43	+8.68	+15.06
Le Monde 94 collection (Clef 2003 campaign)				
Baseline	93.60	88.80	86.13	78.67
$\mathcal{MGB}$-300-3000	95.20	92.00	89.47	81.67
Improvement	+1.70	+3.60	+3.87	+3.81
ATS 94 collection (Clef 2003 campaign)				
Baseline	93.33	91.04	90.56	84.10
$\mathcal{MGB}$-300-3000	96.00	94.60	94.00	86.33
Improvement	+2.86	+3.91	+3.79	+2.65
Le Monde 94 & ATS 94 collections (Clef 2003 campaign)				
Baseline	97.60	96.00	95.07	92.80
$\mathcal{MGB}$-300-3000	98.00	96.40	95.20	93.20
Improvement	+0.40	+0.41	+0.13	+0.43

our tests—*i.e.*, those composed of only IERs, only IARs, and all $\mathcal{MGB}$ rules—give approximately the same improvement. This is justified by:

- INIST is a scientific collection where terms have very weak distributions and marginally co-occur.
- An important part of the vocabulary is not used, since it is not correctly analyzed, due to the used tagger which does not identify specific and scientific terms of INIST.

As illustrated in Table 10, our query expansion approach based on association rules leads to an increase in the exact precision at low recall ($P@5$, $P@10$, $P@15$ and $P@30$ documents) for all collections. This means an increase in the number of retrieved relevant documents put in the head of the top ranked documents list. For example, considering the collection composed by Le Monde 94 & ATS 94, the exact precision at 5 documents is 97.60% for the baseline, which is already a high result. Interestingly enough, even in this situation, our approach based on irredundant association rules offers an improvement equal to 0.40%.

Moreover, we notice that the improvement of the average precision is less significant for high support values. Extracting association rules, when considering a high support value, leads to some trivial associations between terms that are very frequent in the document collection. Therefore, if we expand queries using these terms, we will improve neither the recall nor the precision. For example, in the OFIL document collection, the term *conflict* occurs in the premise of 260 valid association rules. Consequently, the query is expanded by all terms that are in the conclusions of these association rules. For instance, *conflict* has been associated to the following corresponding French words such as: *difficulty*, *solution*, *Bosnia*, *security*, *Serbia*, etc.

5.2 Obtained improvements using generic association rules versus Apriori rules

In Table 11, the improvement obtained using two Apriori rules (Agrawal and Skirant 1994) based approaches (Haddad et al. 2000; Tangpong and Rungsawang 2000) of

Table 11 Improvement of the average precision at 11 points of recall (P@11) for the OFIL and INIST document collections obtained using the proposed approach based on $\mathcal{MGB}$ rules (given in bold) compared to those based on Apriori rules

	minsupp (in number of documents)	*maxsupp* (in number of documents)	*minconf* (in %)	Number of generated rules	Improvement (in %)
OFIL					
Approach of Tangpong and Rungsawang (2000)	1,212	3,305	80	5	+3.06
Approach of Haddad et al. (2000)	110	Not used	20	12,774	+2.58
$\mathcal{MGB}$ rules	5	50	50	5,761	**+29.11**
INIST					
Approach of Tangpong and Rungsawang (2000)	1,633	14,698	5	26	+1.99
Approach of Haddad et al. (2000)	1,655	Not used	10	93,941	+3.52
$\mathcal{MGB}$ rules	3	30	50	3,062	**+12.92**

the literature are compared to ours on both document collections OFIL and INIST. These approaches were in fact not applied on the other document collections we used in our work. The two approaches used the SMART system (Salton 1971) based on $tf \times idf$ weighting scheme and the vector space retrieval model. We thus compare the obtained improvement using our approach to those offered by the aforementioned ones using a common scheme which is based on the $tf \times idf$ weighting.

In Tangpong and Rungsawang (2000), the authors used a high minimum support threshold values which leads to generating few association rules. While in their approach (Haddad et al. 2000), the authors used only terms related to two grammatical categories: the common nouns and the proper nouns. The used rules in this latter approach are composed by a unique term in the premise and a unique term in the conclusion. They then mainly convey relations between a couple of single terms and not between a couple of termsets as it is the case in our approach. The *minsupp* value is set to approximately 1% for both collections that generated 12,774 valid association rules in the case of OFIL and 93,941 in the case of INIST among them a large part is redundant.

In this situation, it is important to not that $\mathcal{MGB}$ rules are not restricted to a given set of terms in the premise or conclusion part, what leads to a more flexible approach. In addition, although the proposed approach in our work uses a lower *minsupp* value (5 documents in the case of OFIL and 3 documents in the case of INIST), the number of generated rules is dramatically lower than that of the approach proposed in Haddad et al. (2000). Note that the used *minconf* value is equal to 50% for mining $\mathcal{MGB}$ rules which is greater than the thresholds used in Haddad et al. (2000). The main aim of using 50% as threshold is to retain strong rules while allowing an interesting number of approximate rules to be mined. These latter rules convey relationships between terms which do not always simultaneously occur. This proves that not only $\mathcal{MGB}$ offers a small set of irredundant association rules but also it ensures very interesting information retrieval performances compared to approaches based on Apriori rules.

5.3 Significance testing

As explained in Smucker et al. (2007), a significance test consists of the following essential elements, namely:

1. A test statistic or criterion by which two approaches are compared—the baseline and the query expansion approach in our case. The IR researchers commonly use the difference in mean average precision (MAP).
2. A distribution of the statistical test given our null hypothesis. A typical null hypothesis is that there is no difference between both compared approaches.
3. A significance level known as the *p-value*. When the significance level is low, we can reject the null hypothesis.

To check whether or not the proposed method of query expansion significantly improves the baseline,[13] we perform the Wilcoxon signed rank test which was used to test pair-wise difference (Smucker et al. 2007). The reason for choosing the Wilcoxon

[13] By baseline, we refer to the baselines using $tf \times idf$, BM25*tf* and Okapi BM25 weighting schemes.

Table 12 Results of Wilcoxon signed rank test ($\alpha = 5\%$)

Runs	MAP	*p-value*
$tf \times idf$ weighting scheme		
OFIL-$\mathcal{MGB}$-5-50	0.3039	0.1060
INIST-$\mathcal{MGB}$-3-30	0.1492	0.3190
Le Monde-$\mathcal{MGB}$-300-3000	0.4104	0.0010
ATS-300-3000	0.5277	0.4300
ATS&Le Monde-$\mathcal{MGB}$-300-3000	0.4441	0.0310
BM25 *tf* weighting scheme		
OFIL-$\mathcal{MGB}$-5-50	0.3497	0.0010
INIST-$\mathcal{MGB}$-3-30	0.1629	0.0438
Le Monde-$\mathcal{MGB}$-300-3000	0.4267	<0.0001
ATS-300-3000	0.5599	0.0395
ATS&Le Monde-$\mathcal{MGB}$-300-3000	0.4778	0.0053
Okapi BM25 weighting scheme		
OFIL-$\mathcal{MGB}$-5-50	0.3536	0.0696
INIST-$\mathcal{MGB}$-3-30	0.1666	0.0840
Le Monde-$\mathcal{MGB}$-300-3000	0.4394	0.0060
ATS-300-3000	0.5688	0.0240
ATS&Le Monde-$\mathcal{MGB}$-300-3000	0.4884	<0.0001

signed rank test is that it is more powerful and indicative test as it considers the relative magnitude in addition to the direction of the differences considered.

Table 12 presents experimental results for a significance level $\alpha = 5\%$. These results show that the best performance for each document collection is obtained while using Okapi BM25 as weighting scheme. The *p-values* given by the Wilcoxon test indicate that the gain between the various runs and the baseline precision are significant. Whereas, once applied on the $tf \times idf$ weighting scheme, the Wilcoxon test is significant only for the collection ATS 94 & Le Monde 94 (for *p-value* $= 0.031$) and the collection Le Monde 94 (for *p-value* $= 0.0010$).

In the current study, experiments highlight that the overall improvements of the BM25*tf* and Okapi BM25 weighting models are better than those of the $tf \times idf$ one.

6 Conclusion and ongoing work

In this paper, we proposed a new approach for query expansion based on association rules between sets of terms. These association rules are extracted from the target document collection by means of mining mechanisms which in turn rely on results from FCA field. Thus, our method computes only the frequent closed termsets and organizes them in a semi-lattice, called Iceberg lattice, together with the associated set of generators to each of them. This structure is then used to extract a set of irredundant rules, which constitute the $\mathcal{MGB}$ basis, representing inter-term correlations in a contextual manner. Interestingly enough, the mined rules convey the most interesting correlations amongst terms w.r.t. our target, namely the association rules-based query expansion.

The experimental study was conducted in this paper on real textual collections using three weighting schemes, namely $tf \times idf$, BM25*tf* and Okapi BM25. The results confirmed that the synergy between association rules and query expansion is fruitful. Indeed, results of the study showed an improvement in the performances of the

information retrieval system, in terms of both recall and precision metrics. This was also highlighted by a carried out significance testing using the Wilcoxon test.

Further work includes using the BIR model integrating the confidence of an association rule into the model's ranking function (de Vries and Roelleke 2005). Indeed, currently, when expanding the query with new terms, the query is considered as a new where all associated terms (original query terms and terms used for the expansion) are considered as equal and processed by the same weighting process. We propose to use a different weighting model for the terms added to the query according to the confidence of the used rules in the expansion. We will also investigate whether the combination of the confidence measure with other measures in the weighting process can be of benefit to the query expansion process. Indeed, expanding a query by a very strong rule and expanding it by a weak rule, according to a some criteria based on selected measures (Guillet and Hamilton 2007), do not lead to queries with equal degrees of correspondence with the initial query. This issue will be studied in detail in a next work.

In the respect, we also plan to address new ways of providing conceptual indexing among queries by creating a representation in terms of irredundant association rules derived from concepts. We will use the support and the confidence of interrelationships between the terms in the candidate concept for the expansion process. In our future weighting schema, the idea is that the global frequency of a term in a query is quantified on the basis of both its own frequency and the confidence of the irredundant association rule in which the term appears.

A special attention will be paid to exact association rules having high confident association rules. For example, the rule $San \Rightarrow Francisco$ with a confidence equal to 1 represents the noun phrase "*San Francisco*". This is justified by the fact that terms of these rules co-occur all the time together in the documents. Discovering such rules (exact association rules with confidence equal to 1) in the context of query expansion is less useful than discovering association rules with confidence less than 1, *i.e.*, approximate ones. But considering *San Francisco* as one text unit and if a query contains *San Francisco*, what are the terms related to *San Francisco* that can be added to the query?

Acknowledgements We would like to thank the anonymous reviewers for their helpful comments and suggestions. We are also grateful to the Evaluations and Language resources Distribution Agency (ELDA) which kindly provided us the Le Monde 94 and ATS 94 document collections of the CLEF 2003 campaign.

References

Agrawal, R., & Skirant, R. (1994). Fast algorithms for mining association rules. In *Proceedings of the 20th international conference on very large databases (VLDB 1994)* (pp. 478–499). Santiago, Chile.

Armstrong, W. W. (1974). Dependency structures of database relationships. In *Proceedings of IFIP congress* (pp. 580–583). Geneva, Switzerland.

Ashrafi, M. Z., Taniar, D., & Smith, K. (2007). Redundant association rules reduction techniques. *International Journal Business Intelligence and Data Mining, 1*(2), 29–63.

Balcázar, J. L. (2010). Redundancy, deduction schemes, and minimum-size bases for association rules. *Logical Methods in Computer Science, 6*(2:3), 1–33.

Bastide, Y., Pasquier, N., Taouil, R., Stumme, G., & Lakhal, L. (2000). Mining minimal non-redundant association rules using frequent closed itemsets. In *Proceedings of the 1st international conference on computational logic (DOOD 2000), LNAI* (Vol. 1861, pp. 972–986). London, UK: Springer-Verlag.

Ben Yahia, S., Hamrouni, T., & Mephu Nguifo, E. (2006). Frequent closed itemset based algorithms: A thorough structural and analytical survey. *ACM-SIGKDD Explorations, 8*(1), 93–104.

BenYahia, S., Gasmi, G., & Mephu Nguifo, E. (2009). A new generic basis of factual and implicative association rules. *Intelligent Data Analysis, 13*(4), 633–656.

Bodner, R. C., & Song, F. (1996). Knowledge-based approaches to query expansion in information retrieval. In *Proceedings of the 11th Biennial conference of the Canadian society for computational studies of intelligence on advances in artificial intelligence (AI 1996), LNCS* (Vol. 1081, pp. 146–158). Toronto, Ontario, Canada: Springer-Verlag.

Buckley, C., Salton, G., Allan, J., & Singhal, A. (1994). Automatic query expansion using SMART: TREC-3. In *Proceedings of the 3rd text retrieval conference (TREC 1994)*.

Croft, B., & Yufeng, J. (1994). An association thesaurus for information retrieval. In *Proceedings of the 4th international conference on computer-assisted information retrieval (RIAO 1994)* (pp. 146–161). New York, USA: CID Press.

de Vries, A. P., & Roelleke, T. (2005). Relevance information: A loss of entropy but a gain for IDF? In *Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR 2005)* (pp. 282–289). Salvador, Brazil: ACM Press.

El-Hajj, M., & Zaiane, O. (2005). Finding all frequent patterns starting from the closure. In *Proceedings of the international conference on advanced data mining and applications (ADMA 2005)* (pp. 67–74). Wuhan, China.

Fonseca, B. M., Golgher, P. B., Pôssas, B., Ribeiro-Neto, B. A., & Ziviani, N. (2005). Concept-based interactive query expansion. In *Proceedings of the 14th international conference on information and knowledge management (CIKM 2005)* (pp. 696–703). Bremen, Germany: ACM Press.

Ganter, B., & Wille, R. (1999). *Formal concept analysis*. Springer-Verlag, Heidelberg.

Grefenstette, G. (1992). Use of semantic context to produce term association lists for text retrieval. In *Proceedings of the 15th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR 1992)* (pp. 89–97). Copenhagen, Denmark: ACM Press.

Guillet, F., & Hamilton, H. J. (2007). *Quality measures in data mining*, Vol. 43. Studies in Computational Intelligence, Springer.

Haddad, H., Chevallet, J. P., & Bruandet, M. F. (2000). Relations between terms discovered by association rules (12 pages). In *Proceedings of the workshop on machine learning and textual information access in conjunction with the 4th European conference on principles and practices of knowledge discovery in databases (PKDD 2000)*. Lyon, France.

Haiquan, L., Jinyan, L., Wong, L., Feng, M., & Tan, Y. P. (2005). Relative risk and odds ration: A data mining perspective. In *Proceedings of the 24th ACM SIGACT-SIGMOD-SIGART symposium on principles of database systems (PODS 2005)* (pp. 368–377). Baltimore, Maryland, USA: ACM Press.

Hu, J., Wang, G., Lochovsky, F. H., Sun, J-T., & Chen, Z. (2009). Understanding user's query intent with Wikipedia. In *Proceedings of the 18th international conference on world wide web (WWW 2009)* (pp. 471–480). Madrid, Spain: ACM Press.

Joho, H., Sanderson, M., & Beaulieu, M. (2004). A study of user interaction with a concept-based interactive query expansion support tool. In *Proceedings of the 26th European Conference on Information Retrieval Research (ECIR 2004), LNCS* (Vol. 2997, pp. 42–56). Sunderland, UK: Springer-Verlag.

Jones, K. S., Walker, S., & Robertson, S. E. (2000). A probabilistic model of information retrieval: Development and comparative experiments. *Information Processing and Management, Elsevier, 36*(6), 779–840.

Kryszkiewicz, M. (2002). *Concise representation of frequent patterns and association rules.* Habilitation dissertation, Institute of Computer Science, Warsaw University of Technology, Warsaw, Poland.

Lin, H. C., Wang, L. H., & Chen, S. M. (2008). Query expansion for document retrieval by mining additional query terms. *Information and Management Sciences, 19*(1), 17–30.

Liu, H., Sun, J., & Zhang, H. (2009). *Post-Mining of association rules: Techniques for effective knowledge extraction*. Chapter V: Post-processing for rule reduction using closed set. IGI Global Publisher.

Lucchese, C., Orlando, S., Palmerini, P., Perego, R., & Silvestri, F. (2003). kDCI: A multi-strategy algorithm for mining frequent sets. In *Proceedings of the IEEE ICDM workshop on frequent itemset mining implementations (FIMI 2003). CEUR Workshop Proceedings* (Vol. 90). Melbourne, Florida, USA.

Luxenburger, M. (1991). Implications partielles dans un contexte. *Mathématiques, Informatique et Sciences Humaines, 29*(113), 35–55.

Mitra, M., Singhal, A., & Buckley, C. (1998). Improving automatic query expansion. In *Proceedings of the 21th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR 1998)* (pp. 206–214). Melbourne, Australia: ACM Press.

Pasquier, N., Bastide, Y., Taouil, R., Stumme, G., & Lakhal, L. (2005). Generating a condensed representation for association rules. *Journal of Intelligent Information Systems, 24*(1), 25–60.

Pfaltz, J. L., & Taylor, C. M. (2002). Scientific knowledge discovery through iterative transformation of concept lattices. In *Proceedings of the workshop on discrete applied mathematics in conjunction with the 2nd SIAM international conference on data mining (SDM 2002)* (pp. 65–74). Arlington, Virginia, USA.

Qui, Y., & Frei, H. P. (1993). Concept based query expansion. In *Proceedings of the 16th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR 1993)* (pp. 160–169). Pittsburgh, PA, USA: ACM Press.

Rungsawang, A., Tangpong, A., Laohawee, P., & Khampachua, T. (1999). Novel query expansion technique using Apriori algorithm. In *Proceedings of the 8th Text REtrieval Conference (TREC 1999)*.

Ruthven, I. (2003). Re-examining the potential effectiveness of interactive query expansion. In *Proceedings of the 26th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR 2003)* (pp. 213–220). Toronto, Canada: ACM Press.

Ruthven, I., & Lalmas, M. (2003). A survey on the use of relevance feedback for information access systems. *Knowledge Engineering Review, Cambridge University Press, 18*(2), 95–145.

Salton, G. (1971). *The* SMART *retrieval system: Experiments in automatic document processing*. Prentice-Hall Series in Automatic Computation, Prentice-Hall, NJ, USA.

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

Salton, G., & McGill, M. J. (1983). *Introduction to modern information retrieval*. McGraw-Hill.

Schenkel, R., & Theobald, M. (2005). Relevance feedback for structural query expansion. In *Proceedings of the 4th international workshop of the initiative for the evaluation of XML retrieval (INEX 2005), LNCS* (Vol. 3977, pp. 344–357). Dagstuhl Castle, Germany: Springer-Verlag.

Shi, X., & Yang, C. C. (2007). Mining related queries from web search engine query logs using an improved association rule mining model. *Journal of the American Society for Information Science and Technology, Wiley, 58*(12), 1871–1883.

Smucker, M. D., Allan, J., & Carterette, B. (2007). A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of the 16th international conference on information and knowledge management (CIKM 2007)* (pp. 623–632). Lisboa, Portugal: ACM Press.

Stumme, G., Taouil, R., Bastide, Y., Pasquier, N., & Lakhal, L. (2002). Computing Iceberg concept lattices with Titanic. *Data & Knowledge Engineering, 2*(42), 189–222.

Sun, R., Ong, C., & Chua, T. (2006). Mining dependency relations for query expansion in passage retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR 2006)* (pp. 382–389). Seattle, Washington, USA: ACM Press.

Tangpong, A., & Rungsawang, A. (2000). Applying association rules discovery in query expansion process. In *Proceedings of the 4th world multi-conference on systemics, cybernetics and informatics (SCI 2000).* Orlando, Florida, USA.

Voorhees, E. M. (1993). Using WordNet to disambiguate word senses for text retrieval. In *Proceedings of the 16th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR 1993)* (pp. 171–180). Pittsburgh, PA, USA: ACM Press.

Xu, J., & Croft, W. B. (1996). Query expansion using local and global document analysis. In *Proceedings of the 19th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR 1996)* (pp. 4–11). Zurich, Switzerland: ACM Press.

Yurekli, B., Capan, G., Yilmazel, B., & Yilmazel, O. (2009). Guided navigation using query log mining through query expansion. In *Proceedings of the 3rd international conference on network and system security (NSS 2009). IEEE computer society* (pp. 560–564). Gold Coast, Queensland, Australia.

Zaki, M. J. (2004). Mining non-redundant association rules. *Data Mining and Knowledge Discovery, 9*(3), 223–248.

Zhai, C. (2001). *Notes on the Lemur tfidf model.* Note with Lemur 1.9 documentation. Technical report, School of Computer Science, Computer Science Department, Carnegie Mellon University (CMU), Pittsburgh, PA, USA.