A Semantically-based Lattice Approach for Assessing Patterns in Text Mining Tasks

John Atkinson, Alejandro Figueroa, and Claudio Pérez

Dept. of Computer Sciences, Faculty of Engineering,
Universidad de Concepcion, Chile

Yahoo! Research, Santiago, Chile

atkinson@inf.udec.cl,claudioperezcarcamo@gmail.com, afiguer@yahoo-inc.com

Abstract. In this paper, a new approach to automatically assessing patterns in text mining is proposed. It combines corpus based semantics and *Formal Concept Analysis* in order to deal with semantic and structural properties for concepts discovered in tasks such as generation of association rules. Experiments show the promise of our evaluation method to effectively assess discovered patterns when compared with other state-of-the-art evaluation methods.

Keywords. Text mining, concept lattices, semantic analysis, association rules.

Un enfoque de lattice basado en semántica para evaluar patrones en tareas de minería de textos

Resumen. En este artículo, se propone un nuevo enfoque para la evaluación automática de patrones en minería de textos. Éste combina semántica basada en corpus y *Análisis Formal de Conceptos* con el fin de manejar propiedades estructurales y semánticas para conceptos descubiertos en tareas tales como generación de reglas de asociación. Los experimentos muestran los resultados promisorios de nuestro método para evaluar efectivamente patrones descubiertos cuando se compara con otros mtodos de evaluación de la literatura.

Palabras clave. Minería de textos, lattices conceptuales, análisis semántico, reglas de asociación.

1 Motivation

Despite the effort of the different text mining approaches to measure interestingness, for example, in association rules, there is no evidence that this actually leads to valuable patterns. For text mining [8, 7] patterns also need to be assessed in terms of interestingness or novelty [1]. However, due to the nature of the natural language texts, this task poses several problems regarding with text analysis. Hence traditional interestingness metrics cannot be applied.

Text Mining can potentially benefit from successful evaluation techniques from Data Mining and Web mining [7]. However, data mining methods cannot be immediately applied to text data for text mining as they assume a structure in the source data which is not present in free text. Hence the assessment of the patterns discovered from texts has almost been a neglected topic in the majority of the text mining approaches.

A major problem is that current text mining methods do not always discover real interesting patterns (i.e., association rules). This may be partially due to the fact that no domain users are involved in the evaluation, domain knowledge is not usually considered and further semantic relationships measuring interestingness are very rare. Patterns evaluation for text mining has been a neglected topic so the few existing approaches fall into two groups: those which are domain-independent [3, 1] and those which use external resources such as lexicons and ontologies. One of the drawbacks with the first one is that it does not take advantage of domain models in order to assess implicit relationships within the generated patterns so as to assess their novelty. Whereas, the second kind of approach is highly dependent on the existence and organization of electronic concept resources such as ontologies and thesaurus.

Accordingly, in this paper, a new approach which combines structural and semantic features is proposed in order to accurately assess the real interestingness of patterns discovered (i.e., association rules) by a traditional text mining task. The model uses structural knowledge extracted

Computación y Sistemas Vol. 17 No. 4, 2013 pp. 467 -476
ISSN 1405-5546
DOI:10.13053/CyS-17-4-2013-001

from domain model by using Formal Concept Analysis (FCA) methods [3, 4] in order to assess the novelty of patterns and has it augmented with semantically-based knowledge via *Latent Semantic Analysis* (LSA) [1, 5] which is applied to enable partial matching in FCA [4].

The paper is organized as follows: section 2 discusses the main approaches to evaluate interestingness in some specific Text Mining task (i.e., association rules generation), in section 3 the combined model for pattern evaluation which uses FCA and LSA is presented, section 4 describes and discusses the main experiments using the model and state-of-the-art metrics and human experts, and finally section 5 draws the main conclusions of the work.

2 Patterns Evaluation in Text Mining

A major problem with state-of-the-art evaluation metrics to assess discovered patterns in data mining is that they all have been designed to deal with structured relational data (i.e., support, confidence, surprisingness, etc). Instead, in text mining the data are unstructured and so they cannot be easily interpreted by computers. The lack of structure raises the difficulty of uncovering the implicit knowledge inside the documents. Furthermore, there is a huge amount of involved linguistic implicit and explicit knowledge (i.e., lexical, syntactical, semantic, etc) which makes it very difficult to identify what a pattern should look like and to evaluate its degree of novelty. Thus, there are plenty of well-established text mining approaches but just a few approaches concerned with real knowledge discovery from texts in which finding novel stuff is a key issue.

A promising early approach which indeed dealt with quantitative pattern evaluation for text mining measured the degree of novelty of rules discovered from texts extracted from Web collections based on existing lexical and semantic information in the general-purpose lexical database $WordNet$. The evaluation involves assessing the coverage of a rule (i.e., number of items covered by the rule in a training set), and the semantic interestingness. For this, items of the rule's antecedent and consequent are evaluated according to the semantic distance between them in $WordNet$. The working assumption here is that the longer the semantic distance is, the more novel the relation is, and therefore, the rule. Resulting experiments show that the system evaluation somewhat correlates with human judgments.

Nevertheless, this evidence shows that a discovery task which depends on general-purpose conceptual resource may produce misleading results because of the lack of domain-specific knowledge: unconnected terms–longer distances–may lead to interesting patterns even if they do not exist in the knowledge base [5, 8, 10].

Overall, the notion of novelty or interestingness of these approaches relies on a specific organization of conceptual resources, whether they are domain-independent (e.g., $WordNet$) or specific-domain (e.g., UMLS in the medical domain). However, the effectiveness of the methods is affected in terms of robustness as the discovered knowledge is highly dependent on the existing information, and the particular semantic acquisition task in mind.

A more robust strategy which builds knowledge models from scratch, measures the interestingness degree of association rules [3] based on the distance between antecedent and consequent of the rule based on a concept lattice built by using *Formal Concept Analysis* (FCA) methods [4]. FCA is a theory of data analysis that identifies conceptual structures among data sets and produces graphical visualizations of the inherent structures among data that can be understood as knowledge model (i.e., ontology). It is an exploratory method for data analysis and provides non-trivial information about input data of two basic types-concept lattice and attribute implications. A concept is a cluster of similar objects (similarity is based on presence of same attribute values); concepts are hierarchically organized (specific vs. general).

For this FCA-based rule evaluation approach, a lattice hierarchy is used to compute conceptual distance between concepts. A concept is usually composed of two parts: *extension* and *intention*. Extension covers all the objects belonging to the concept whereas the intention covers all the objects' valid attributes for a concept. In FCA, a triple (G, M, I) is called a $context$ where G and M are sets and $I \subseteq G \times M$. The elements of G and M are called $objects$ and $attributes$ respectively. The context is often represented by a cross-table of objects (i.e., documents) versus attributes (i.e., terms or words).

For this approach, the knowledge model (K) represents generalization relationships between concepts composed of terms extracted from a text corpus. For any concept $A, B \in K$, $A \subseteq B$ iff any instance of A is also an instance of B. Formally, the lattice $(K, \subseteq)$ is a directed graph containing K vertices, and the relation $\subseteq$ defines the edges of the graph [3].

Using this concept lattice, the degree of interestingness of a rule is measured in terms of the probability of going from one term k_1 to other term k_2 contained in a concept of the lattice model k what is called $conformity$, with these terms appearing in the antecedent and consequent part of the rule. Formally, the $conformity$ of a rule $k_1 \rightarrow k_2$ according to model $(K, \subseteq)$ is the probability of transition for finding a path from k_1 to k_2 in the model $(K, \subseteq)$. A connection between terms is regarded to as $uninteresting$ as long as this is a direct 'translation' of the relationship between K_1 and K_2 hence longer paths are preferred. For example, if the term "fruit" is more general than "apple" (apple $\subseteq$ fruit), then the rule "apple $\Rightarrow$ fruit" will have a high $conformity$ and therefore it becomes 'uninteresting'.

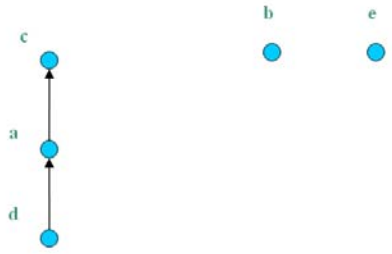

Fig. 1. A simplified Concept Lattice

This strategy uses the $ACLOSE$ algorithm to generate association rules [8, 7]. Using these rules, the created knowledge model is used to determine the interestingness degree of each rule. A sample model to assess pattern is seen in figure 1, where each edge between nodes (i.e., concepts) represents a generalization relationship (i.e., k_1 is a set of k_2). For the example in which terms are extracted from text documents, $conformity$ is then computed keeping two kinds of relations in mind: that between "a" and "c" where there exists a path between the nodes (figure 1), and that between "c" and "d" where there is no path at all connecting them. Here, the $conformity$ value for $a - c$ is greater than for $c - d$, so the generalization relationship between "a" and "c" is stronger than for "c" and "d", hence the $c - d$ rule may express a more interesting pattern. This is mainly due to the existence of a model that connects the nodes "a" and "c'", whereas for the $c - d$, no path that connects them was even found.

A main drawback of this evaluation method is that human supervision is still required to build and adjust the domain model when necessary. On the other hand, creating concept lattices strongly relies on exact matching of inclusion relationships between objects. Hence only exact attributes are considered when creating a concept in the lattice which may discard a significant number of approximately $similar$ concepts [5].

3 A Multi-Strategy Approach to Patterns Evaluation

One of the key issues with patterns evaluation when building knowledge models (i.e., lattices) is its restricted kinds of semantic relationships. This can be partially due to the way the concepts are created. In the case of lattices, creating concepts strongly depends on exact properties of set inclusion which may discard semantically similar terms which are not in the set built for the concept. On the other hand, purely semantic distance evaluation based on LSA does not allow us to establish specific generalization/specialization relationships which makes difficult to compute structure-based metrics such as $conformity$ or $novelty$.

For this, our work combines lattice-based models and LSA-based similarity metrics in order to provide approximate matching using semantic distances of sets when creating knowledge models and it, in turn, enables a more effective mechanism for assessing patterns in the form of association rules. Thus, this research contributes a new multi-strategy approach to assess patterns using automatically created semantically-based knowledge models as follows:

— *Providing a multidimensional semantic space:* contextual meaning of words in a corpus can be inferred from occurrences across text documents via semantic analysis. Multidimensional vector obtained from this task can then be used to compute semantic

similarity/distance between terms when creating domain models.

— *Providing knowledge models by building lattices:* discovered patterns can be assessed by measuring distances in the underlying hierarchical concept structure in terms of generalization/specialization connections [3]. Unlike other approaches, lattice construction is extended to include LSA-based semantically similar attributes when creating concepts which in turn, may make rule evaluation more effective as the concept creation becomes more robust.

From now on, this two-phase strategy which uses corpus-based semantics is called a *Semantic Concept Lattice* (SCL), which allows for the approximate creation of concepts in a lattice-like knowledge model. Thus, our approach is also capable of determining implicit similarity relationships which cannot be usually captured by structural methods such as FCA (i.e., synonyms). The overall procedure can be seen in algorithm 1 which requires preprocessing, knowledge models generation tasks before evaluating association rules.

Algorithm 1. Building semantic lattices and evaluating patterns

1: Let T a given input training text corpus
2: Let P a set of association rules to be assessed
3: $T' \leftarrow$ Text Preprocessing using T
4: $S \leftarrow$ Creating Semantic Spaces (i.e., vectors) with LSA for T'
5: $L \leftarrow$ Generating the Semantic Concept Lattice using T' and S
6: Evaluate P according to the $Conformity$ measure on L

3.1 Generating Association Rules

Using simple terms extracted from a text corpus, frequent closed itemsets and association rules are automatically generated by using the $ACLOSE$ and $APRIORI$ algorithms respectively [7]. Usual terms included combinations of $Noun - Adjective$, $Noun - ProperName$. It is important to highlight that while complex linguistic structures might be extracted from the text to represent the rules, lattice creation restricts its inclusion sets to only contain terms.

3.2 Text Preprocessing

In this stage, some basic preprocessing tasks for handling natural language texts were carried out including: *Stopwords removal* (i.e., non-relevant words are removed from the text corpus in order to avoid inferences of highly frequent words when computing LSA spaces), *Lemmatization* (i.e., involves the reduction of the words in a corpus to their respective lexemes), *Part-of-Speech (POS) tagging* (i.e., words in a corpus are marked as corresponding to a particular part of speech or lexical category), *Features Extraction* (i.e., relevant terms are looked for in order to get a representative set of features representing the documents).

3.3 Generating Knowledge Models

Semantic and domain models are created based on LSA and FCA, in order to build our *Semantic Concept Lattice* (SCL):

— *Creating Semantic Spaces with LSA:*

Latent Semantic Analysis (LSA) is a kind of mathematical technique that generates a high-dimensional semantic space (aka. a set of semantic vectors) from the analysis of a huge text corpus. Specifically, words, sentences or paragraphs can be represented by these semantic vectors. The ultimate goal of LSA is to find a data mapping which provides information well beyond the lexical level and reveals semantical relations between the entities of interest [5].

This latent structure is obtained by extracting and inferring relations of expected contextual usage of words in passages of texts. A first step represents the text as a matrix in which each row stands for a unique word and each column stands for a text passage or other unit. Each cell contains the frequency with which the word of its row appears in the passage denoted by its column. Next, the cell entries are subjected to a preliminary transformation in which each cell frequency is weighted by a function that expresses both the word's importance in the particular passage and how much information the word type carries in the domain of discourse. Next, LSA applies Singular Value Decomposition (SVD) methods [5] to the matrix so that reconstructed matrices in fewer dimensions are said to capture the

latent structure of the terms co-occurring in texts.

By keeping track of the patterns of occurrences of words in their corresponding contexts, LSA is able to recover the latent structure of the meaning space, this is, the relationship between meanings of words: the larger and the more consistent their overlap, the closer the meanings.

Specifically, by using a text corpus, LSA produces a set of highly-dimensional semantic vectors representing knowledge at the lexicosemantic level for each term of a document. These terms will become part of the concepts when creating a lattice. Vectors are obtained by a combination of matrix-based operations on the occurrences of terms and based on SVD. This vector representation can then be used to measure a semantic closeness between two term vectors t_1 and t_2 in a semantic space S as follows:

$$Similarity\_LSA(t_1, t_2, S) = cosine(\vec{t_1}, \vec{t_2}) \quad (1)$$

— *Generating the Semantic Concept Lattice:*

In order to create our *Semantic Concept Lattice* (SCL), concepts are built from terms extracted from a corpus, for which an adaptation of FCA methods was used [3]. Our approach considers lexicosemantic knowledge provided by LSA in order to add close terms into a concept of a lattice without using any domain resource as seen in algorithm 2. Note that a lattice uses a kind of hierarchical relationship (i.e., $a - subset - of$) which cannot be determined by pure LSA as semantic closeness is symmetrical. Thus it is not possible to determine the named relationship or even its direction. In order to deal with this, LSA semantic vectors for each terms forming a concept, are introduced in order to measure the similarity between terms being added by FCA when creating the concepts.

In the algorithm, $NEIGHBORS$ computes the upper neighbors of a concept (G, M) in which G becomes the set of objects (i.e., documents) and M becomes the set of attributes (i.e., terms or features). It can be used to recursively compute all the concepts in L of a context by starting from the smallest concept $(\emptyset', \emptyset'')$ of the lattice. Every concept c has two lists associated with it: the list of c^* of its upper neighbors and the list of c^* of its lower neighbors.

One object may be shared by two different concepts as their upper neighbor. While the algorithm processes each of the two concepts their shared upper neighbor must be detected in order to get the relationships right. For this purpose all concepts are stored in L. Each time the algorithm finds a neighbor, it looks for it ($lookup(..)$) in L to find previously inserted instances of that concept. In case the concept is found, the existing lists of neighbors are updated. Otherwise, if the LSA similarity on the semantic space S between the new concept and x exceeds some threshold, the previously unknown concept is entered into the lattice.

The algorithm inserts concepts into L and looks them up at the same time: $next(c, L)$ asks for the smallest concept that is greater than c with respect to the total order $\prec$. To make sure all concepts that are inserted are also considered for their upper neighbors,

Algorithm 2. Creating a *Semantic Concept Lattice* (SCL)

CREATE_SCL(T',S):
with T' and S representing (G, M) and the LSA semantic vectors respectively:

1: Let L be LATTICE (G, M, I) where (G, M) and (G, M, I) are the concept and context respectively
2: $c \leftarrow (\emptyset', \emptyset'')$
3: insert(c, L)
4: **for all** x ∈ $NEIGHBORS\ (c, (G, M, I))$ **do**
5: $lookup(x, L)$
6: **if** NotFound and $Similarity\_LSA(x, c, S) > Threshold\_LSA$ **then**
7: insert (x, L)
8: **end if**
9: $x_* \leftarrow x_* \cup \{c\}$
10: $c^* \leftarrow c^* \cup \{x\}$
11: $c \leftarrow next\,(c, L)$
12: **if** NotFound **then**
13: exit
14: **end if**
15: **end for**
16: SCL ← L

the total order $\prec$ must relate to the partial lattice order $\leq$ in the following way: $c_1 < c_2$ implies $c_1 \prec c_2$. This way, recently inserted neighbors are greater than the actual concept with respect to $\prec$ and will be considered later by $next(..)$.

Thus, given two terms (t_1 and t_2) to be incorporated into a lattice, the term t_1 may be included in a subset of t2 (with t_1 being not necessarily an exact matching of the other term) only if the LSA similarity between both exceeds certain threshold ($threshold_{LSA}$). This lattice so created is called a *Semantic Concept Lattice* (SCL).

Thus, if the semantic threshold ($threshold_{LSA}$) was maximum (1.0), it would mean that a very flat and uninteresting structure will be created (see first lattice of figure 2) as $conformity$ will be uniform for all the terms of the hierarchy.

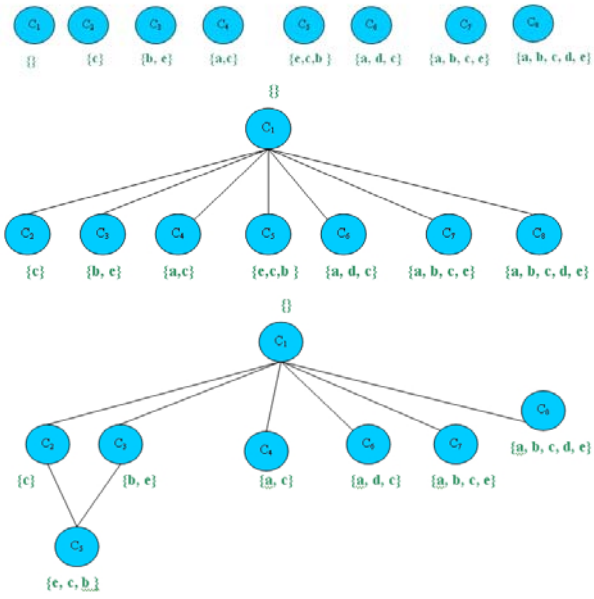

Fig. 2. *SCL* for $threshold_{LSA}$ values of 1.0, 0.9 and 0.8 respectively

When reducing the $threshold_{LSA}$ down to 0.9 and 0.8, second and third lattice of figure 2 are obtained respectively in which distances between the terms are much clearer so that if they appear in some association rules their $conformity$ would be lower, and therefore they become potentially more interesting. On the other hand, for $threshold_{LSA}$ values under 0.6, the generated concept structure tends to be the same as lattice creation methods which do not use semantic distances [3]. This may be due to the fact all the input terms are candidates to be part of a concept so the inclusion relation becomes a major issue.

3.4 Evaluating Association Rules

The interestingness degree of each discovered rule is assessed by using the automatically created SCL. For this, each rule is measured in terms of its $conformity$ value which is based on two basic tasks:

1. *Computing Conformity:*

   For each term in the SCL, $conformity$ is computed against the rest of the terms of the lattice. $Conformity$ represents the degree of generalization/specialization of two terms in a concept structure so that the longer the distance between them, the lower their $conformity$ is, and therefore, the relationship becomes interesting.

2. *Generating a Conformity Ranking:*

   Each assessed rule according to its semantic conformity is ranked by starting from the most interesting one (lower $conformity$) to the less interesting one (higher $conformity$). Thus effectiveness of an evaluation can be seen as the position of each rule on this ranking.

4 Experiments and Results

We assessed the effectiveness of our patterns evaluation approach by building a computer prototype and carrying out a series of adjusting and final experiments. Final results were compared with some state-of-the-art metrics for patterns evaluation and they were then correlated with human judgment.

Note that our patterns evaluation approach is domain-independent but for experimentation purposes we used a corpus of nearly $100,000$ documents extracted from the Web and the *Corpus BioText Data*[1]. Natural Language Processing tools such as Part-of-Speech (POS) taggers and stemmers were obtained from *GENIA* and *SNOWBALL*[2], respectively. In order to

[1] http://biotext.berkeley.edu/data.html
[2] http://snowball.tartarus.org/

create our *Semantic Concept Lattice* (SCL), the tool *Concepts*[3] was applied and extended. Furthermore, LSA-based vectors were obtained using the *Infomap*[4] library.

4.1 Setting Experiments

For preprocessing and training purposes, a $k-fold$ ($k = 10$) cross-validation method was used for documents extracted from the original corpus so as to set and adjust parameters for our model which included $threshold_{LSA}$, depth of the lattice, etc. In addition, values of traditional metrics such as support/confidence (i.e., *minsupport* and *minconfidence*) were systematically looked for so that ACLOSE and APRIORI algorithms were applied to generate relevant association rules [7].

By manually examining association rules generated by state-of-the-art algorithms, the most relevant patterns were obtained for *minsupport* and *minconfidence* values under 1.0 and 0.1, respectively. Furthermore, $threshold_{LSA}$ values were set by analyzing the quality of the knowledge lattice being created in terms of the $conformity$ of the discovered rules.

Experiments show that for threshold values between 0.0 and 0.3, the SCL becomes irrelevant as it only contains one hierarchical level, so that the assessed rules will be given a uniform $conformity$. The average depth of the SCL built for different threshold values showed that for a threshold value of 0.4, the highest depth of the lattice was obtained (approx. 3). In this scenario, the most interesting rule seems to be *"segments"* $\rightarrow$ *"transmembrane, Orientation"* which is assessed in the first place. On the other hand, rules such as *"protein"* $\rightarrow$ *"Arf, GTPase-activating"* are evaluated in the last position of the ranking hence it becomes 'uninteresting'. For $threshold_{LSA}$ values of 0.5, two different positions are generated in the ranking in the first position for two rules respectively: *"segments"* $\rightarrow$ *"transmembrane, Orientation"* and *"fur"* $\rightarrow$ *"Max-Planck-Institut, Zuchtungsforschung"*. On the contrary, rules *"protein"* $\rightarrow$ *"Arf, GTPase-activating"* and *"two-hybrid analysis"* $\rightarrow$ *"cerevisiae, screening"* are evaluated in the last place.

For a different set of generated rules, table 1 shows rankings for $threshold_{LSA}$ values under 0.2 which clearly affects the evaluation ranking of the model. For example, for $threshold_{LSA}$ value of 0.2, the rule *"crassa"* $\rightarrow$ *"Neurospora"* is evaluated in the first place whereas the rules *"virus"* $\rightarrow$ *"Vaccinia"* and *"actin"* $\rightarrow$ *"cerevisiae"* are located in the last position of the ranking. Hence better relevant rules are obtained from the SCL for $threshold_{LSA}$ of 0.2, in terms of a deeper level in the structure and different positions in the ranking.

Table 1. Ranking for different rules in the SCL for increasing values of $threshold_{LSA} < 0.2$

	Threshold		
Rule	**0.0**	**0.1**	**0.2**
profilin $\rightarrow$ cerevisiae	3	3	4
actin $\rightarrow$ cerevisiae	3	3	5
glucan $\rightarrow$ Golgi	4	4	2
homolog $\rightarrow$ Drosophila	2	2	3
virus $\rightarrow$ Vaccinia	1	3	5
crassa $\rightarrow$ Neurospora	4	1	1
genome $\rightarrow$ Arabidopsis	4	1	4

4.2 Final Experiments

In order to evaluate the effectiveness of the proposed approach, previous settings were used to adjust the model and compare it against other state-of-the-art association rules evaluation metrics. For this, $k - fold$ cross-validation was used with the original corpus being the testing set. Furthermore, obtained evaluations were correlated with human judgment so as to investigate the effectiveness of the interestingness automatic evaluation.

Association rules were generated from a combination of 104 *features* on POS tags such as *proper name*, *name* and *adjective*, given a final set of 153 different terms. A set of 25 random rules were finally obtained and the $threshold_{LSA}$ value was set to 0.4 based on previous setting experiments.

Final evaluation and comparisons were carried out by using the following state-of-the-art evaluation metrics:

M1: uses typical metrics such as *support* and *confidence*.

M2: uses the $conformity$ metric based on simple lattices [3].

[3] http://www.st.cs.uni-sb.de/~lindig/src/concepts.html
[4] http://infomap-nlp.sourceforge.net/

Table 2. Ranking association rules using different evaluation metrics

	Method				Expert	
Rule	$M1$	$M2$	$M3$	$M4$	$E1$	$E2$
aggregation → Huntington, disease-associated	4	3	23	1	1	2
nuclear → cerevisiae, pore	1	2	12	1	2	2
protein → wild-type	4	2	13	1	2	3
multidrug → Pdr5p, ABC	4	2	10	1	1	1
archaeal → jannaschii, aIF6	4	3	7	1	1	1
substrate → amino	3	1	24	2	3	3
mammalian → C, Class	4	2	25	2	3	1
amino→substrate	3	2	11	2	5	5
Two-hybrid analysis, system→cerevisiae	4	2	4	2	2	5
cerevisiae→Saccharomyces	4	2	5	2	2	2
stalk→cerevisiae, ribosomal	4	3	6	2	1	1
cells→parental, Chemical	4	2	8	2	2	2
motif→Walker, nucleotide-binding	4	2	9	2	5	5
fungus→mushroom-producing, Schizophyllum	4	3	12	2	5	1
pore→cerevisiae, nuclear	4	2	12	2	1	1
genetic data→ DSC1/MBF	4	2	14	2	5	5
stalk→cerevisiae, ribosomal	4	2	16	2	4	3
high→cerevisiae, mobility	4	2	17	2	3	2
system→cerevisiae, Two-hybrid analysis	4	2	18	2	4	5
intermediate→covalent, Biochemical	4	3	19	2	5	4
jannaschii→archaeal, aIF6	4	3	20	2	1	4
domain→LexA, DNA-binding	4	2	22	2	5	5
albican→Candida, strains	4	3	1	2	3	4
box→A, ATP-binding	4	2	2	2	5	5
bacterium→Gram-negative, Escherichia	4	1	3	3	1	1

M3: uses the semantic dissimilarity between antecedent and consequent of a rule based on information provided by pure LSA with no external or domain models.

M4: uses our SCL method to measure the $conformity$ of the generated rules.

At the same time, each rule was also assessed by two human experts (*E1* and *E2*) who measured the degree of interestingness of the rules in the same scale as the models. However, for clarity's sake, it was 'normalized' to a scale ranging from 1 (`very interesting`) to 5 (`uninteresting`). Evaluation made by the four automatic methods and the humans can be seen in table 2.

The table shows that evaluation method (metrics) $M1$ assesses in the first position the rule *"archaeal"* → *"jannaschii, aIF6"*, whereas the same rule is seen as very interesting by both experts. It may be due to the fact that metrics $M1$ uses only statistical-based metric to assess rules which may be not very useful to assess interestingness/novelty. On the other hand, metrics $M2$ generates three different evaluations, having in the first place the two rules *"bacterium"* → *"gram-negative, escherichia"* and *"substrate"* → *"amino"*, whereas the experts assessed 7 rules in a first place (out of 9 different rules).

Metrics $M3$ evaluates in a first place the rule *"albican"* → *"candida, strains"*. However, the same rule is regarded to as not that interesting by both experts (range 3 and 4). Rules evaluated using our metrics ($M4$) are within the three different positions of the ranking which suggests several coincidences in some places. Hence our model seems to better evaluate (i.e., higher positions) those rules that might be more interesting. Since there is no absolute real notion of the interestingness degree, evaluation and ranking produced by the four metric were correlated with human experts for the same

set of rules. For this, the *Spearman* correlation (r) was computed for all the assessments as seen in table 3.

In general, a promising and positive correlation was observed between experts and our metrics as compared with the rest of the evaluation methods ($p < 0.01, t = 3.461$). Furthermore, the pure lattice-based evaluation method showed almost no association with the real assessment. Our approach ($M4$) shows a fair predictive ability for evaluating patterns, although one expert was more demanding in the evaluation than the other one. This may mainly be due to that $E1$ has more than 20 years of experience so he got more background knowledge to assess the patterns.

On the other hand, the pure LSA-based metrics ($M3$) is better correlated with experts than the lattice-based approach ($M2$) but worse than our evaluation method ($M4$). A lower correlation of metrics $M2$ can also be due to the fact that the metrics considers only structural relationships, so that the when there are implicit semantic connections between terms, the metrics fails to detect them.

Table 3. Correlation between experts and evaluation metrics

	Metrics			
Expert	$M1$	$M2$	$M3$	$M4$
$E1$	0,01	-0,05	0,01	0,32
$E2$	-0,02	-0,04	0,07	0,25

5 Conclusions

In this paper, a new metrics that combines lattice-based methods and corpus based semantics is proposed to automatically assess the interestingness degree of association rules. It allows the approach to filter and effectively rank patterns finally delivered to users making decisions. The model uses LSA to allow for an approximate matching of set inclusion when building concept structures in the form of lattices. It makes the approach more robust to include implicit semantic relationship which does have not necessary generalization/specialization links. Our strategy builds a *Semantic Concept Lattice* in order to measure $conformity$ of association rules. Assessment of interestingness using SCL was well correlated with human judgment as compared with other evaluation methods. Furthermore, the model benefits from its resource-independent and semantically-based nature which can make it domain independent.

Overall, results suggest that our evaluation approach is indeed effective to assess the interestingness degree of simple discovered patterns in text mining in comparison with other evaluation methods and human performance. In addition, combining lattice-based creation methods and corpus based semantics is very promising to assess discovered patterns as compared with both methods separately (LSA and lattice-based methods).

Acknowledgements

This research was partially supported by FONDEF (Chile), under grant number CA12i10081: *"Text Routing: Reconocimiento Inteligente de las Intenciones de Usuarios para Enrutamiento de llamadas en Call Centers utilizando Técnicas de Análisis Semntico de Lenguaje Natural y Métodos de Aprendizaje Automático"*.

References

1. **Atkinson, J. & Rivas, A.** (**2008**). Discovering novel causal patterns from biomedical natural-language texts using bayesian nets. *IEEE Transactions on Information Technology in Biomedicine*, 12(6), 714–722.
2. **Bie, T. D.** (**2011**). An information theoretic framework for data mining. *Proc. of the 17th ACM SIGKDD conference on Knowledge Discovery and Data Mining (KDD'11), San Diego*.
3. **Cherfi, H., Napoli, A., & Toussaint, Y.** (**2004**). Knowledge-based selection of association rules for text mining. *16th European Conference on Artificial Intelligence - ECAI'04. (Valencia, Spain)*, 24, 485–489.
4. **Fu, H., Jennings, B., & Malone, P.** (**2007**). Analysis and representation of biomedical data with concept lattice. *IEEE/IES Conference on Digital Ecosystems and Technologies, Cairns Australia*.
5. **Landauer, T., McNamara, D., Dennis, S., & Kintsch, W.** (**2007**). *Handbook of Latent Semantic Analysis*. Lawrence Erlbaum Associates.

6. **Mampaey, M., Tatti, N., & Vreeken, J.** (**2011**). Tell me what i need to know: succinctly summarizing data with itemsets. *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '11)*.
7. **Olson, D. & Delen, D.** (**2008**). *Advanced Data Mining Techniques*. Springer.
8. **Srivastava, A. & Sahami, M.** (**2009**). *Text Mining: Classification, Clustering, and Applications*. Chapman and Hall/CRC Publishers.
9. **Tatti, N. & Vreeken, J.** (**2011**). Comparing apples and oranges: measuring differences between data mining results. *Proceedings of the 2011 European conference on Machine learning and knowledge discovery in databases - Volume Part III (ECML PKDD'11)*.
10. **Yuefeng, L., Abdulmohsen, A., & Ning, Z.** (**2010**). Mining positive and negative patterns for relevance feature discovery. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '10. ACM, New York, NY, USA, 753–762.

John Atkinson is a full professor of the Department of Computer Sciences at Universidad de Concepcion, Concepcion, Chile. He received his Ph.D. degree in Artificial Intelligence from the University of Edinburgh, Scotland, UK (2003). His current research interests include natural language processing, text mining, and evolutionary computation. Dr. Atkinson is a member of the AAAI, IEEE and senior member of the ACM.

Alejandro Figueroa is a researcher at *Yahoo! Research*, Santiago, Chile. He received his Ph.D in Computational Linguistics from Universitat des Saarlandes, Saarbrucken, Germany (2010). Dr. Figueroa has also been a researcher at DFKI (the German Center for Artificial Intelligence), Saarbrucken, Germany, and Yahoo! Research Lab in Barcelona, Spain. His research interests question-answering systems, natural language processing, machine learning and information retrieval.

Claudio Pérez is an internet project manager working for a private company in Concepcion, Chile. He received his MSc in Computer Sciences from Universidad de Concepcion, Chile (2010). Mr. Pérez has also been involved in educational technology and text mining projects. His main interests include text mining, internet technology and management, and business intelligence.

Article received on 12/07/2013; accepted on 25/09/2013.