Available online at www.sciencedirect.com
ScienceDirect
Computers & Graphics 31 (2007) 316–326
COMPUTERS & GRAPHICS
www.elsevier.com/locate/cag

# Visual text mining using association rules

A.A. Lopes*, R. Pinho, F.V. Paulovich, R. Minghim

*Universidade de São Paulo, Instituto de Ciências Matemáticas e de Computação CP 668, São Carlos 13560-970, São Paulo, Brazil*

**Abstract**

In many situations, individuals or groups of individuals are faced with the need to examine sets of documents to achieve understanding of their structure and to locate relevant information. In that context, this paper presents a framework for visual text mining to support exploration of both general structure and relevant topics within a textual document collection. Our approach starts by building a visualization from the text data set. On top of that, a novel technique is presented that generates and filters association rules to detect and display topics from a group of documents. Results have shown a very consistent match between topics extracted using this approach to those actually present in the data set.


*Keywords:* Visual text mining; Association rules; Data mining; Information visualization

## 1. Introduction

As a convenient way for humans to store information, text documents hold a significant amount of the information available to institutions, corporations and individuals. In fact, about 85% of all information found in businesses is exclusively stored in some kind of unstructured data [1]. These might be spread among letters, reports, claims, presentations, news articles, e-mail messages, intranet pages and services, web pages and other types of electronic documents.

In order to deal with this kind of data, a lot of effort has been devoted to the research of text mining techniques. As defined by Tan [2], text mining relates to the process of extracting interesting and non-trivial patterns or knowledge from unstructured textual documents. The unstructured nature of textual information is the main challenge here and adds difficulty to the already complex data mining tasks. Text mining is also referred to as text data mining or knowledge discovery from textual databases.

Among the diverse set of techniques for data mining, association rules have been extensively used to represent knowledge [3] and have also been used in text mining [4–6]. Association rules for textual documents code important information on relationships between them that ultimately are capable to identify patterns, themes and context.

Due to the complex tasks involved in the mining of text data sets, information visualization techniques should play a central role in increasing the performance of the text analysis process. The broader synergy between data mining and visualization is further explored with visual data mining techniques (VDM) [7], which integrates the process of visualization and mining in order to insert the user in the process of building a more adequate mental model of a particular data set. In VDM for texts, or visual text mining (VTM), multi-disciplinary approaches are gathered to allow users to understand general structure and local trends in complex sets of documents.

Chen [8] suggests that navigating through a visual information space, as may be the case with VTM, requires the user to build an internal cognitive map, as when navigating through the real world. Thus, it is desired to have visualizations where the user is capable of establishing a connection with his or her cognitive map, while avoiding the inherent complexity of the underlying information space. This mapping of information and cognitive

*Corresponding author. Tel.: +55 16 3373 9678; fax: +55 16 3373 9751.
*E-mail addresses:* alneu@icmc.usp.br (A.A. Lopes), rpinho@icmc.usp.br (R. Pinho), paulovic@icmc.usp.br (F.V. Paulovich), rminghim@icmc.usp.br (R. Minghim).

0097-8493/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.cag.2007.01.023

navigational spaces is evident on some techniques of knowledge domain visualization (KDViz). A clear example is the use of cartographic maps to visualize thousands of conference abstracts [9].

In this paper, we propose a framework of tools for VTM that integrate multi-dimensional visualization and exploration techniques with mining techniques (particularly Association Rules) to extract meaning from a text data set. With visualization, a content-based map of documents is built, from which a large number of exploration tasks can be carried out. We also present a novel technique to extract topics and subjects from groups of documents belonging to that map by means of adequately generating and filtering association rules for the documents. The integration of visualization and exploration tools with automatic detection of topics yields very meaningful results, as it will be shown. The proposed approach helps the user to achieve understanding of the knowledge domain represented in a particular collection of documents, as well as examine the individual subjects approached in parts of that collection.

The next section presents related work on topic extraction and theme based visualizations. Section 3 briefly presents the visualization technique and software developed to visualize text data sets and then focus on the method developed to extract themes from association rules (Section 3.3). Section 4 presents various maps built with the application of these techniques and discusses the contributions presented here.

## 2. Related work

Topic Maps is an ISO standard for representing knowledge linked to information resources (web pages, electronic documents, etc.) [10]. They convey a semantic network that allows for navigating on higher abstraction levels than that of individual resources [11]. With this notation, complex structures may be described by the simple concepts of: (i) topics, (ii) occurrences (resources), and (iii) topic associations. Visualization techniques have been proposed for exploring topic maps [12,13].

However, topic information for a given resource or set of resources is not always available. Some information retrieval approaches provide to users descriptions of subsets of search results, that could be interpreted as a topic or theme for each subset. These descriptions can be derived from previously defined categories and classification algorithms. For example, Chen and Dumais [14] apply support vector machines to assign web search results to a given set of classes. Another approach is to consider some external data model as description, that could be, for example, a terminology model [15].

More related to our proposal is the use of unsupervised techniques, such as clustering, for topic finding. On the Scatter/Gather system [16], search results are clustered and summaries (descriptions) are built as a list of frequent terms in each cluster. The authors have shown that the system performed better for an information retrieval task than the simple presentation of ranked results.

With a similar purpose, another technique selects phrases (*n-grams*) from clusters and use them as descriptions [17]. Phrases are ranked according to a set of properties. Each phrase is the seed for a cluster that contains all documents holding that *n-gram*. Clusters with high overlapping, above 75%, are merged. Finally, clusters are presented according to the rank of phrases. Document ranking inside each cluster could reflect either its rank for the original query or its similarity to the seed phrase for the cluster. This approach is similar to the Grouper system [18], and can be viewed as an evolution of it. Grouper has a simpler method for phrase ranking and displays multiple phrases for each cluster.

A document map technique by Skupin [19] assigns labels to hierarchical clusters over documents on a map built using a self-organizing map (SOM) [20]. For high-level clusters, the term with the highest count is chosen. For lower level clusters, each cluster is viewed as a document and the label is built from the three highest ranking terms according to a non-specified variation of the term frequency, inverse term frequency weight [21].

ThemeScape is a topic driven visualization where a landscape is built by successively layering the computed contributions (weights) of thematic terms over a ground plane where documents were distributed according to their similarity [22]. If a given document has a term, the height of its region is increased proportionally to term contribution. Term weighting is based on their ability to discriminate previously derived clusters. The weight $w_k$ for a term $k$ in a cluster $i$ is given by

$$w_k = \frac{f_{k,i}}{\sum_{j \neq i} f_{k,j}},$$

where $f_{k,l}$ is the frequency of term $k$ on cluster $l$.

Chen [8] extracts and labels 'specialties' for a citation landscape. The top three components or factors are computed from a co-citation matrix by principal component analysis (PCA). Each factor depicts a specialty and is named after the most frequent term found on the titles of top 10 articles related to that factor. Each document is then shown with an associated red, green and blue bar, where the length of each color represents the relevance of each specialty for that particular document.

ThemeRiver is designed to show thematic changes over time for a collection of documents where each document is time tagged [23]. Each theme is shown as a colored river that widens or straitens according to the strength of the theme for the documents from a particular time period. Theme strength is given by the same strategy applied in SPIRE, and cited above in the description of Themescape [22].

TopicIsland is a technique that displays theme changes and also subparts on specific topics found in a single document, derived from a wavelet transform analysis [24].

Our approach differs from the above because it is based on association rules extraction. The motivation for adopting this alternative over previous methods is the fact that association rules carry an intrinsic semantic level not possible to determine from simple frequency counts or from dimension reduction. It achieves that by compounding individual term contributions according to their co-occurrence in a text. As a consequence they can be extended to consider various levels of abstractions due to its variable degree of generalization. Therefore, it is our expectation that our approach can be extended to support automatic detection of contexts within a text data set. None of the other approaches lend themselves to this kind of investigation.

## 3. Mining document collections using multi-dimensional projection and rules

The process to build an interactive document map consists of three steps. First, any necessary pre-processing of the texts is performed, such as term extraction and counting for building a vector representation of the data set. Following that a multi-dimensional projection is performed to position the documents as points in 2D space and construct a triangulation or graph from the data. That can be done either from the coordinates of the vector space or from a similarity measure between texts. The final map can be interacted with in a number of ways to provide exploration of relationships between texts and groups of texts. The final step described here is the detection of themes by properly generating and selecting association rules from the documents contents.

### 3.1. Multi-dimensional projections for document collections

In this section the process to create a map from a collection of unstructured data is described. For text, such process is composed by two main tasks: (1) create a vector representation of the document collection, whereby each document is represented as a vector on a $\mathbb{R}^n$ space; and (2) perform a projection of this representation to a $\mathbb{R}^2$ space in a way that allows visualization and exploration of inherent structures within the document collection.

In order to create the vector representation, first the frequencies of terms ($n$-grams) in each document are counted, ignoring the 'stopwords', i.e. non-representative words, such as prepositions, articles, etc. After that, a Luhn's cut-off [25] is performed to eliminate terms that are either too frequent or too rare, since they do not contribute to establish distinction amongst texts. This process tries to select the most relevant terms of the collection. Finally, each document is represented as a vector where the dimensions are represented by the set of remaining terms, and the coordinates are the frequency count of such terms in each document, weighted according to Term-frequency Inverse-document-frequency (TfIdf) [21].

The second step to create the document map employs a *projection technique* (PT) in order to map each vector representing a document to a point on a plane. Here we follow the definition of PT given by Tejada et al. [26]. According to such definition, if $X$ is a set of points in $\mathbb{R}^n$ with $d : \mathbb{R}^n, \mathbb{R}^n \to \mathbb{R}$ a criterion of proximity between points in $\mathbb{R}^n$, and $P$ is a set of points in $\mathbb{R}^2$ with $\hat{d} : \mathbb{R}^2, \mathbb{R}^2 \to \mathbb{R}$ a proximity criterion in $\mathbb{R}^2$. A PT can be described as a function $\alpha : X \to P$ which tries to approximate $|d(x_i, x_j) - \hat{d}(\alpha(x_i), \alpha(x_j))|$ as close as possible to zero, $\forall x_i, x_j \in X$.

In this paper, we employ two main projection techniques developed by members of our research team: *projection by clustering* (ProjClus) [27]; and *least-square projection* (LSP) [28].

The former aims at projecting the data using local relations in $\mathbb{R}^n$, where $|d(x_i, x_j) - \hat{d}(\alpha(x_i), \alpha(x_j))|$ must be as close as possible to zero, $\forall x_i \in X$, but with $x_j \in S_i$, where $S_i$ is a set of points belonging to a neighborhood of $x_i$. This technique first creates $\sqrt{n}$ clusters, then projects the centroid of each cluster on the plane using the *force scheme* (FS) [26]; in the subsequent step the technique individually projects the points of each cluster to $\mathbb{R}^2$ using FS. Finally the cluster's elements are arranged in their final layout according to the projection of its centroids.

The second technique also takes into account local relationships in $\mathbb{R}^n$, but it is based on a least-square technique originally employed in surface modeling and reconstruction [29]. In such a technique, first a reduced subset of points in $X$, called 'control points', is projected onto $\mathbb{R}^2$ by a high-precision projection method, such as FS. After that, making use of a neighborhood relationship of the points in $\mathbb{R}^n$ and the cartesian coordinates of the control points in $\mathbb{R}^2$ it is possible to build a linear system whose solution is the projection of all the points in $\mathbb{R}^2$.

Since these techniques are based on local relations between points in $\mathbb{R}^n$, both are suitable to project document vector representations to $\mathbb{R}^2$. The space created in these representations are very sparse and high dimensional. In such cases, the points are normally arranged along local subspaces and they are related only with a small number of nearest neighbors inside the same subspace, thus working outside of their neighborhood may cause distortions on the final layout [30].

Fig. 1 shows two document maps generated from *IEEE* 2004 *contest* data set [31], using ProjClus (Fig. 1(a)) and LSP (Fig. 1(b)) techniques. In such maps the points are colored by the frequency of the word 'graphs' on the documents. Red points have no occurence of the word.

### 3.2. Projection Explorer (PEx)

In order to support the creation and interactive exploration of document maps, we developed a tool, called *Projection Explorer* (*PEx*),[1] where LSP, ProjClus, and the remaining techniques presented here are implemented. In this tool the user inputs a document collection in ASCII

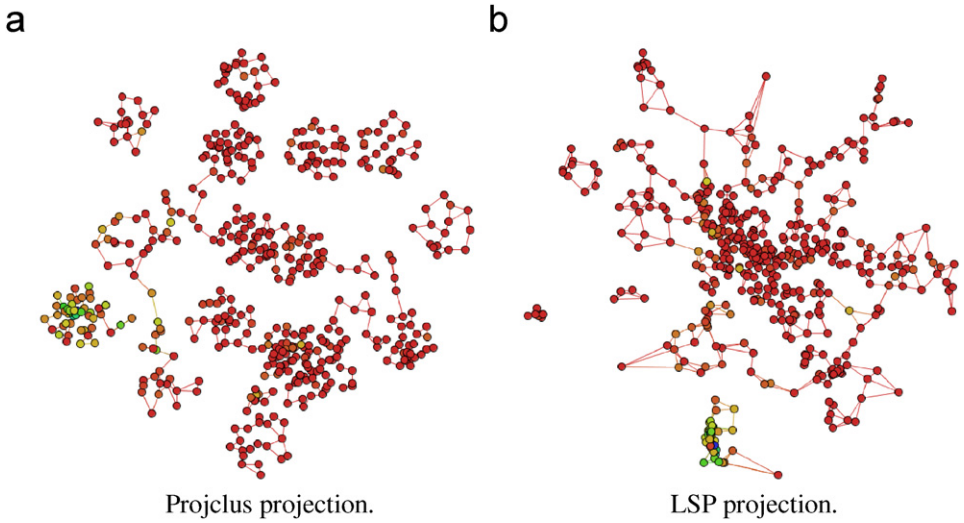


Fig. 1. Two different document maps for the IEEE 2004 Contest data set: (a) ProjClus projection; (b) LSP projection.

format (it is also possible to execute a query on the Web and use the results as a collection), and set the parameters to create the document vector representation and the parameters of the chosen projection technique. PEx, therefore, supports the full process of document map creation and exploration. Fig. 2 shows the main window of the PEx program.

The major features of PEx to explore document maps include: (1) the possibility to find the nearest neighbors of a document using the distances on the final projection and in the multi-dimensional original space ($d$, as defined in the previous section); (2) coloring the documents (points) on the map according to the frequency of a word or group of words; (3) label creation to identify a group of documents on the map—these labels are based on the most frequent words which occurs on a group of documents selected on the projection; (4) view of the content of a document or group of documents, as well as their neighbors by simply double clicking over a point on the map; (5) coordinating two different projections, i.e. selecting some documents on a projection and see where this documents are placed on another projection; (6) connection and exploration of the points according to a triangulation or a neighboring relationship (in either $\mathbb{R}^2$ or $\mathbb{R}^n$) and (7) coloring of the documents according to their distance to a selected document.

In order to use many of the features available in the PEx program, the user must select a limited area on the document map. This selection area is a rectangular portion of the projection and the selected points can be described as a set of points $S_k = \{p \in P | \min\{r_{x1}, r_{x2}\} \leqslant p_x \leqslant \max\{r_{x1}, r_{x2}\}, \min\{r_{y1}, r_{y2}\} \leqslant p_y \leqslant \max\{r_{y1}, r_{y2}\}\}$, where $p_x$ and $p_y$ are the coordinates of the point $p$, and $r_{x1}$, $r_{x2}$, $r_{y1}$ and $r_{y2}$ are the coordinates of the two points which defines the selection area.

The following text describes the use of association rules in the exploration of a text map, presenting the technique devised to extract themes from those rules in regions of the map under user selection.

### 3.3. Topic extraction using association rules

The following text is a formal statement of the problem of mining association rules from text. Let $I = i_1, i_2, \ldots, i_m$ be a set of literals or items (representing a term from the bag of words). Let $D$ be a set of transactions, where each transaction $T$ is a set of items such that $T \subseteq I$ (representing a document). An association rule (AR) is an implication of the form $X \Rightarrow Y$, where $X \subset I$, $Y \subset I$, and $X \cap Y = \emptyset$. The rule $X \Rightarrow Y$ holds in the document set $D$ with confidence $c$ if $c\%$ of the documents in $D$ that contain $X$ also contain $Y$. The rule $X \Rightarrow Y$ has support $s\%$ in $D$ if $s\%$ of transactions in $D$ contain $X \cup Y$.

The term to the left of the arrow is called head and the set of remaining terms is called the body of the rule.

Intuitively, an appropriate set of AR should be capable of describing a subject in which that term appears as central or a topic related to a set of documents. For instance, given the set of association rules from a case-based reasoning corpus (title, authors, abstract and references from 250 articles), one (who knows CBR field) can realize that these rules are related to the well-known article 'CBR: Foundational Issues, Methodological Variations, and System Approaches' by Aamodt and Plaza [32]:

aamodt ← combine foundational issues general ($s = 3.3$, $c = 100.0$)

aamodt ← custom variations plaza application ($s = 3.3$, $c = 100.0$)

aamodt ← environment foundational plaza architecture ($s = 3.3$, $c = 100.0$).

[1]The *Projection Explorer* (*PEx*) tool and its manual are available at http://infoserver.lcad.icmc.usp.br

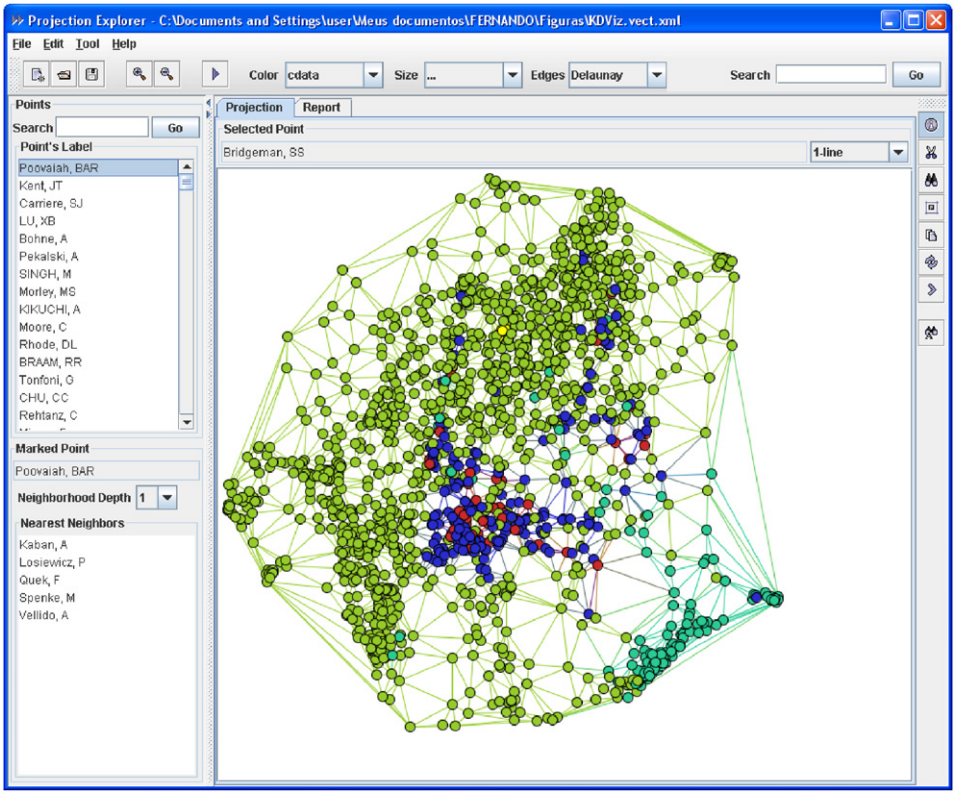


Fig. 2. Projection Explorer (PEx) main window.

Further abstracting the rules' composition it is possible to identify different contexts within a text set.

In other words, good association rules can be seen as good candidates for describing a set of documents or define a topic or theme to be assigned for the set. A major problem with this approach is how to deal with the hundreds of thousands of rules produced, and the consequent redundancy of rules and irrelevant terms that appear in those rules.

The algorithm proposed here extends the Apriori algorithm by Agrawal and Srikant for interactively and iteratively generating association rules from a selected set of documents [33]. Three major changes are implemented in order to overcome the above mentioned shortcomings.

In order to deal with the problem of irrelevant attributes, we define a measure for rules that takes into consideration the relevance of the terms in the rule, besides its confidence and support. The relevance is higher for a term or combination of terms that maximally represents or describes a subset of documents (cluster) and minimally describes others (the remaining documents of the corpus). In this sense, we define the weight of a term as a relative frequency of the term inside a selection (cluster) of documents by its frequency in the entire corpus. This weight carries the traditional meaning of the TfIdf. Intuitively, terms with higher frequency in the selection and not frequent in the entire corpus are more likely to be a better candidate for relevant term to describe the content of the selection. The weight of a term is given by

$$W_{t_j S_k} = \frac{\sum TfIdf_{t_j S_k}}{\sum TfIdf_{t_j C}},$$

where the weight of the term $t_j$ in the selection $S_k$ is the summation of the *TfIdf* of the term in $S_k$ divided by the summation of the *TfIdf* of the term in the entire corpus $C$. As the *Idf* is a constant, we can write

$$W_{t_j S_k} = \frac{\sum Tf_{t_j S_k}}{\sum Tf_{t_j C}}.$$

Our experiments have shown that a poor selection of attributes (bag of words) impairs substantially the 'quality' of the rules. This occurs mainly because frequent terms, which usually are not relevant for discriminating documents in the data set, tend to appear in most rules. The generation of association rules with terms with higher weight $W_{t_j S_k}$ avoids or minimizes these non-interesting rules.

We deal with the second problem, the large amount of rules, as follows. Instead of post-pruning the set of rules generated from all documents in $D$ by some rule quality measure, we generate rules which contain at least one term (seed) present in a selected set $T$ of most weighted terms. The number of seeds is given by an ad hoc value that we have set to 10 as a result of numerous experiments.

Additionally, we have initially set the minimum support for the Apriori Call in Algorithm 1 at 50%. This (relatively) high initial support also limits the number of rules produced by Apriori, restricting them to those that are representative of the selection. Since not every selection is consistent enough to have rules with this support, we iteratively lower the minimum support requirement until at least one rule is found. Eventually, no rule is found even with a very low minimum support and the process ends.

**Algorithm 1.** Iterative generation and ranking of association rules.

Input: Selection $S_k$ % $k$ selected points (document × term matrix ($M_{K,m}$)
Corpus $C$ % $n$ points, matrix ($M_{N,m}$)
Bag_of_words % $m$ terms of matrix document × term
Output: ranked association rules subset (SR)

For each term $t_j$ from the Bag_of_words
  $Tf_{jC}$ (total frequency of the term $j$ in the corpus $C$)
  $Tf_{jS}$ (total frequency of the term $j$ in the selection $S$)
  $W_{t_jS_k} = \frac{Tf_{jS_k}}{Tft_{jC}}$ (relative frequency)
  $s(t_j \leftarrow)$ in $S_k$ (support of each 1-itemset)

$Minsup = 50\%$
$S = \emptyset$ % rule subset
$T = \emptyset$ % $n$ most weighted terms of Bag_of_words (seeds)
$1\_itemsets = \emptyset$
$2\_itemsets = \emptyset$
$n = 10$ % number of seeds

Do
  For each $t_j \in S_k$
    $1\_itemsets = 1\_itemsets \cup \{t_j \in Bag\_of\_words | Sup(t_j \leftarrow) > MinSup\}$
    $W_{t_{jS},t_{jC}} = \frac{\sum Tf_{t_j,S_k}}{\sum Tft_j,C}$
  %Select $n$ seeds ($n$ terms $t_j$ from 1_itemsets with the larger $W_{j_S,j_C}$)
  $T = \{t_1, \ldots, t_n\}$
  $2\_itemsets = 2\_itemsets \cup \{(t_i, t_j), i \neq j | (t_i, t_j) \in T \times 1\_itemsets\}$
  Call Apriori for producing rules with 2 ⟨number of literals⟩ $m$
  If $S = \emptyset$
    $MinSup = 0.75 * MinSup$
    $n = n + 1$
While ($S = \emptyset$ and $MinSup >= 0.01$)

$SR = ranking(S, n)$ % $n$ most ranked rules

Return $SR$

Finally, for each frequent item set, a variable number of rules can be induced by choosing different elements from this set as head. To avoid redundancy, we create only one rule for each item set, with one of the terms as the head. So, for instance, for an item set with 5 terms, 5 different options exist. All of these rules have the same support, but their confidence may vary. The option effectively chosen is one amongst those with highest confidence.

Labels can be shown for any number of the top ranked rules, according to the sum of the weight of their terms (Algorithm 2). In our examples, we have chosen to display only the top ranked rule. Algorithms 1 and 2 synthesize the iterative process for generation and ranking of candidate rules.

**Algorithm 2.** Ranking.

Input: $S, N$ % association rules set, Number of Rules
$S_k$ % $k$ selected points (document × term matrix ($M_{K,L}$)
Corpus $C$ % $n$ points, matrix ($M_{N,L}$)
Bag_of_words % $L$ terms of matrix document × term
Output: $N$ top ranked rules subset ($SR$ - selected rules) which covers all documents ($d_i$)
$SR = \emptyset$
For each $AR_i = (t_1 \leftarrow t_2, t_3, \ldots, t_m) \in S$
  $w_{AR_i} = \sum_{j=1}^{m} w_{t_j,S}$
  $conf(AR_i)$ % confidence of rule $AR_i$

$Sort(AR, w_{AR})$
$SR = \{AR_1, \ldots, AR_N\}$
Return $SR$

## 4. Results

In this section we present some results of our technique by demonstrating the exploration of two different corpora. The first experiment was carried out in a corpus named CBR-ILP-IR. This corpus is composed by 574 scientific articles (with title, abstract, and references) in three different subjects: case-based reasoning (CBR), inductive logic programming (ILP), and information retrieval (IR). The CBR and ILP articles were obtained manually from the Lecture Notes on Artificial Intelligence (LNAI) series, and the ILP ones were retrieved from the Web. The second experiment employs another corpus, named NEWS, which is composed by 2684 RSS news feed articles, collected from BBC, Reuters, CNN, and Associated Press sites, during two days in April 2006.

The goal of these experiments is to show that multi-dimensional projections and association rules can be successfully integrated to: (i) visually identify similarity relations amongst documents in the collection, (ii) help the user to achieve a deeper understanding of subjects in a selected region of the map, and (iii) explore the relationships amongst documents in the selected area and in its possible sub-areas.

Fig. 3 presents a map created using the ProjClus technique. It also presents the association rules generated by choosing five different regions on the map, each region corresponding to the five most distinct groups of points, determined visually. In the picture, the rule for region 1 indicate documents related to communication, protocols,

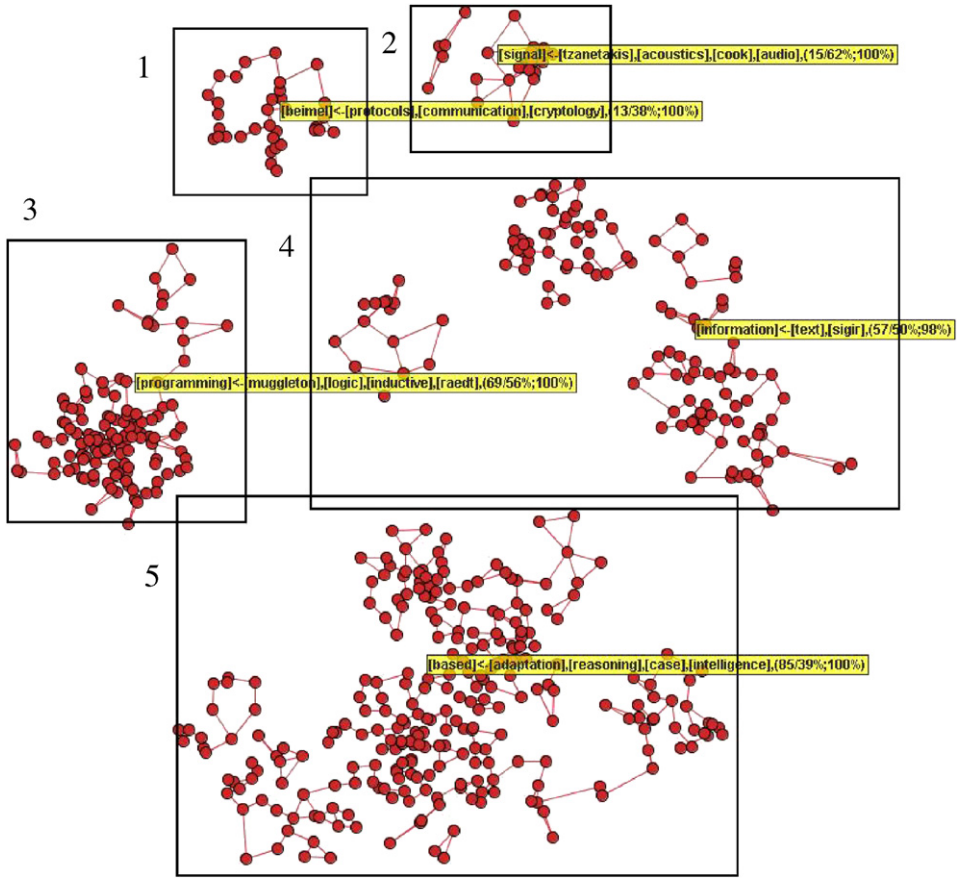


Fig. 3. Overview CBR-ILP-IR map with rule labels. Numbers in parentheses on labels are number of covered documents, support and confidence over selected documents, respectively. Label 1: [beimel] ← [protocols], [communication], [cryptology] (13/35%;100%). Label 2: [signal] ← [tzanetakis], [acoustics], [cook], [audio] (15/62%;100%). Label 3: [programming] ← [muggleton], [logic], [inductive], [raedt] (69/56%;100%). Label 4: [information] ← [text], [sigir] (57/50%;98%). Label 5: [based] ← [adaptation], [case], [intelligence] (85/39%;100%).

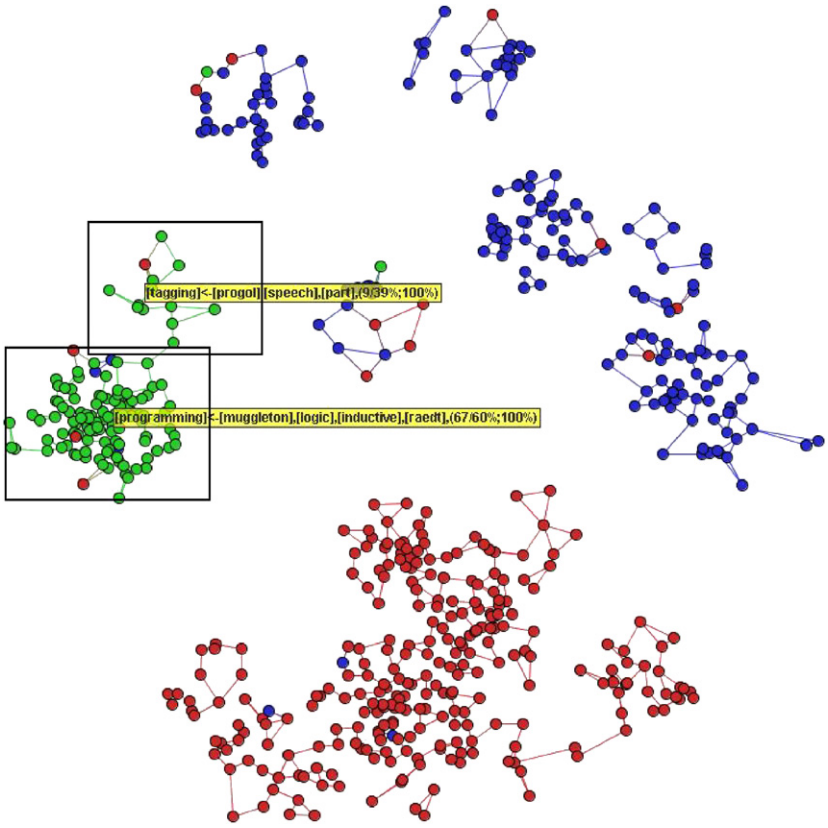


Fig. 4. CBR-ILP-IR colored by manual classification. Here the two major groups of documents inside the ILP subject are tagged. Top label: [tagging] ← [progol], [speech], [part] (9/39%;100%). Bottom label: [programming] ← [muggleton], [logic], [inductive], [raedt] (67/%;100%).

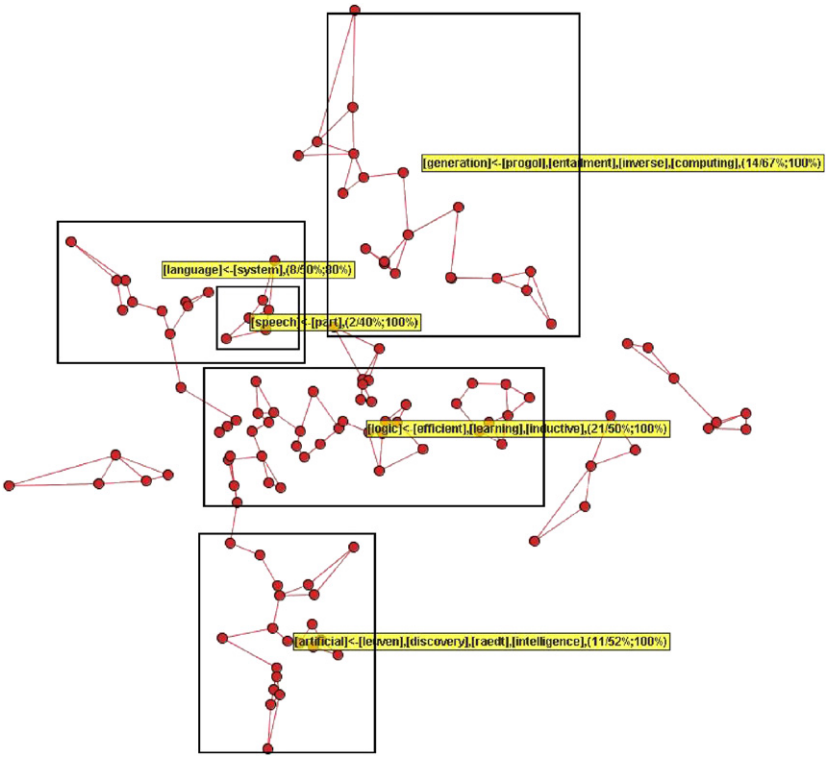


Fig. 5. A deeper view of the ILP subject. Labels (top to bottom): [generation] ←[progol], [entailment], [inverse], [computing] (14/67%;100%), [language] ← [system] (8/50%;80%), [speech] ← [part] (2/40%;100%), [logic] ← [efficient], [learning], [inductive] (21/50%;100%), [artificial] ← [leuven], [discovery], [raedt], [intelligence] (11/52%;100%).

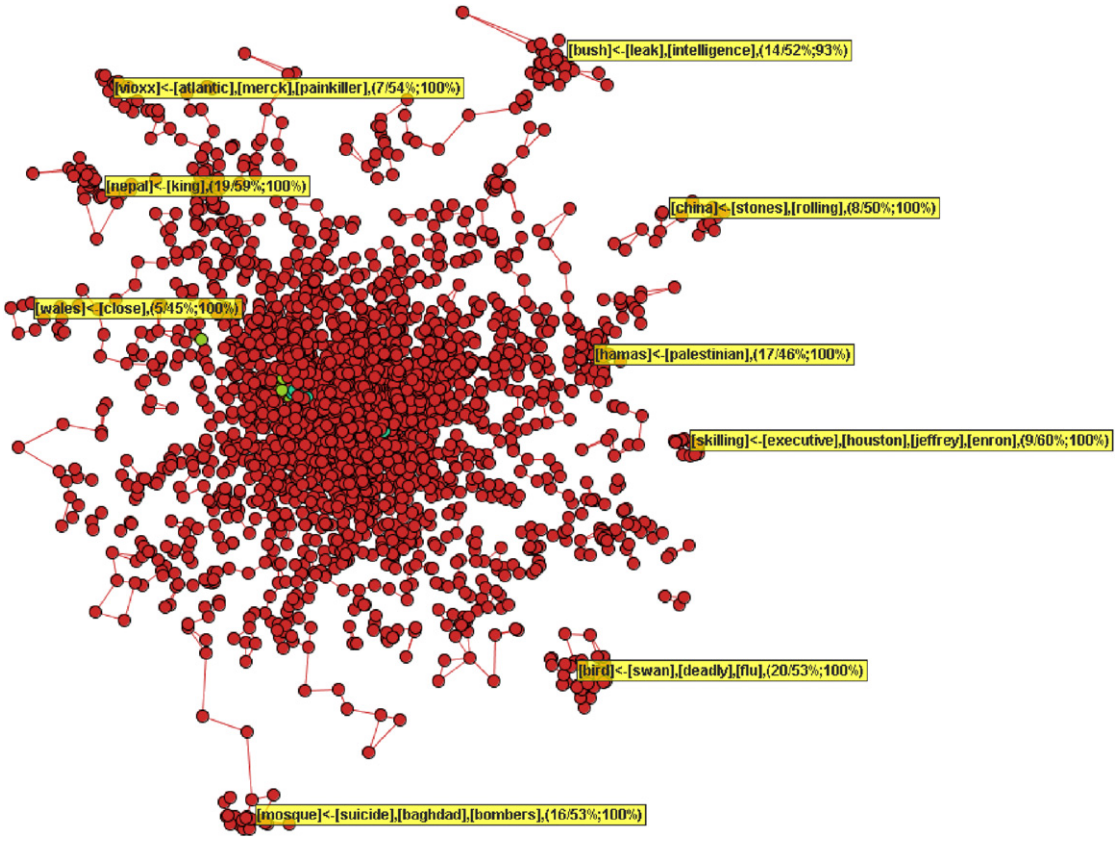


Fig. 6. NEWS Corpus map. Labels (clockwise, from topmost): [bush] ← [leak], [intelligence] (14/52%;93%), [china] ← [stones], [rolling] (8/50%;100%), [hamas] ← [palestinian] (17/46%;100%), [skiling] ← [executive], [houston], [jeffrey], [enron] (9/60%;100%), [bird] ← [swan], [deadly], [flu] (20/53%;100%), [mosque] ← [suicide], [baghdad], [bombers] (16/53%;100%), [wales] ← [close] (5/45%;100%), [nepal] ← [king] (19/59%;100%), [vioxx] ← [atlantic], [merck], [painkiller] (7/54%;100%).

cryptology, and their occurrences imply, with 100% of confidence, that the term 'beimel' (an author in the field) appears. Rules for regions 2 and 4 are related to the IR field, split into audio and text retrieval, respectively. For region 3, the rule indicate documents of the ILP field. Finally, the rule for region 5 indicate documents on CBR and adaptation process in CBR.

Fig. 4 shows the same map presented in Fig. 3, now colored by a classification determined manually based on the source of the articles.[2] Looking at the ILP region, one can see two groups, one which is related to Part-of-speech tagging and a larger one related to ILP in general as well as inductive learning.

In order to explore a group of documents further we can select that group from the map and export its points as a new corpus. Fig. 5 shows a new map generated from exported points of the ILP region of the previous map, in which a group arises that is described by the rule (language ← system) and, within that, the group related to Part-of-speech (Pos) Tagging appears again. Considering the articles in the group, one can see that they are related to inductive learning and language, and specifically to the use of ILP on the task of Pos Tagging. That demonstrated that the techniques hold their capabilities under detailed exploration of parts of a large corpus.

In the second experiment reported here, using the NEWS corpus, we show that it is possible to achieve similar results using a corpus with non-scientific documents. Fig. 6 shows a document map for that corpus, now using the LSP projection technique. Due to the number of documents in this corpus, there is a dense region in the center of the map, and some groups that are better defined, which were placed around it. The user can continue to explore the dense area zooming in it (a PEx feature) or applying the same process used in the previous experiment. The rule tags that appear over the regions on the map correctly identify the central themes approached by the news agencies of the articles those two days (bombings in Iraq, White House leak investigation, bird flu concerns, and so on). These examples identify the capabilities of the proposed techniques.

Next we summarize the conclusions reached from our results.

## 5. Conclusions

Results presented in this paper illustrate that the synergy between visualization and data mining techniques

[2]As such, a certain degree of overlapping between articles of different classes is expected.

can attain promising results in the analysis of complex data.

Our main contribution was devising a novel technique to extract themes from subareas of a document map using association rules, thus taking advantage of the ability of AR to compound individual concepts into a more abstract one. As a consequence, the definition of themes by this approach can be used consistently in various levels of specialization inside subregions of a map, as defined by the user.

Such an approach for text exploration must rely on a display that adequately places documents according to their similarity. For that purpose, we have integrated our topic extraction method with a geometric layout built using fast high precision multi-dimensional projection techniques that have previously demonstrated to be effective in positioning points according to similarity.

The selection of association rules for text that relies solely on support and confidence requires a tedious manual filtering process, which might also prove to be unable to reach a satisfactory result. The term weighting process proposed here for AR selection substantially relieves this problem for the application of theme extraction. It also reduces the need to remove irrelevant terms before rule generation.

One question that often rises when dealing with vector representation for text sets is which size of $n$-gram to use. The approach presented here is capable of, based on 1-grams, producing via association rules some of the relationships that would be coded in other $n$-grams. This avoids the typically enormous size of the bag of words for $n$-grams when $n>1$.

It is essential for the method that the underlying visual map adequately reflects content-based neighborhood relations. Therefore, when projections are not adequately generated, theme detection is not as accurate. On the other hand, theme detection can be employed as an indicator of the quality of the projection.

Besides considering association rules for theme detection as a starting point for semantic evaluation of document maps, we intend to study the expansion of this technique to support automatic identification of distinct contexts within a general text data set.

## Acknowledgments

This work was funded by financial agencies FAPESP, São Paulo, Brazil (process numbers 04/07866-4, 04/09888-5 and 05/02263-2) and CNPq (process number 304758/2005-1), Brasília, Brazil.

## References

[1] Bess C, Lehmann J, Patel B, Schmidt K, Phifer W, Williamson J. The grand challenges of information technology. In: Engineering management conference, 2003. IEMC'03. Managing technologically driven organizations: the human side of innovation and change; 2003, p. 610–5.

[2] Tan A. Text mining: the state of the art and the challenges; 1999.

[3] Hand D, Mannila H, Smyth P. Principles of data mining. Cambridge, MA: The MIT Press; 2001 [Adaptive computation and machine learning].

[4] Cherfi H, Napoli A, Toussaint Y. Towards a text mining methodology using frequent itemsets and association rules. Soft Computing Journal 2004:11.

[5] Kodratoff Y. Knowledge discovery in texts: a definition, and applications. In: Proceedings of the 11th international symposium on foundations of intelligent systems; 1999. p. 16–29.

[6] Rajman M, Besançon R. Text mining: natural language techniques and text mining applications. In: Proceedings of the seventh IFIP 2.6 working conference on database semantics (DS-7); 1997.

[7] Oliveira MCF, Levkowitz H. From visual data exploration to visual data mining: a survey. Visualization and Computer Graphics. IEEE Transactions 2003;9(3):378–94.

[8] Chen C. Information visualization: beyond the horizon. Berlin: Springer; 2004.

[9] Skupin A. The world of geography: visualizing a knowledge domain with cartographic means. Proceedings of the National Academy of Sciences of the United States of America 2004;101:5274–8.

[10] International Organisation for Standardization (ISO) and International Electrotechnical Commission (IEC). Topic maps. International Standard ISO/IEC 13250:1999; 1999.

[11] Le Grand B, Soto M. Visualisation of the semantic web: topic maps visualisation. In: Information visualisation, 2002. Proceedings. Sixth international conference on; 2002, p. 344–9.

[12] Le Grand B, Soto M. Topic maps et navigation intelligente sur le web sémantique. In: AS CNRS Web Sémantique; 2002.

[13] Baudon O, Auillans P, Jarry F. Using xml-topic map on a pda. In: XML conference & exposition; 2001, p. 1.

[14] Chen H, Dumais S. Bringing order to the web: automatically categorizing search results. In: CHI '00: Proceedings of the SIGCHI conference on human factors in computing systems. New York, NY, USA: ACM Press; 2000. p. 145–52.

[15] Pratt W. Dynamic organization of search results using the UMLS. American Medical Informatics Association Fall Symposium 1997;480:4.

[16] Hearst MA, Pedersen JO. Reexamining the cluster hypothesis: scatter/gather on retrieval results. In: SIGIR '96: Proceedings of the 19th annual international ACM SIGIR conference on research and development in information retrieval. New York, NY, USA: ACM Press; 1996. p. 76–84.

[17] Zeng HJ, He QC, Chen Z, Ma WY, Ma J. Learning to cluster web search results. In: Proceedings of the 27th annual international conference on research and development in information retrieval; 2004. p. 210–7.

[18] Zamir O, Etzioni O. Grouper: a dynamic clustering interface to web search results. Computer Networks 1999;31(11–16):1361–74.

[19] Skupin A. A cartographic approach to visualizing conference abstracts. IEEE Computer Graphics and Applications 2002;22(1): 50–8.

[20] Kohonen T. The self-organizing map. Proceedings of the IEEE 1990;78(9):1464–80.

[21] Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. Information Processing Management 1988;24(5):513–23.

[22] Wise JA. The ecological approach to text visualization. Journal of the American Society for Information Science 1999;50(13):1224–33.

[23] Havre S, Hetzler E, Whitney P, Nowell L. ThemeRiver: visualizing thematic changes in large document collections. IEEE Transactions on Visualization and Computer Graphics 2002;8(1):9–20.

[24] Miller NE, Wong PC, Brewster M, Foote H. TOPIC ISLANDS TM—a wavelet-based text visualization system. In: Visualization'98. Proceedings; 1998. p. 189–96.

[25] Luhn HP. The automatic creation of literature abstracts. IBM Journal of Research and Development 1958;2:159–65.

[26] Tejada E, Minghim R, Nonato LG. On improved projection techniques to support visual exploration of multi-dimensional data sets. Information Visualization Journal 2003;2(4):218–31.
[27] Paulovich FV, Minghim R. Text map explorer: a tool to create and explore document maps. In: Information visualization (IV06), London: IEEE Computer Society Press; 2006.
[28] Paulovich FV, Nonato LG, Minghim R, Levkowitz H. Visual mapping of text collections through a fast high precision projection technique. In: Information visualization (IV06), London: IEEE Computer Society Press; 2006.
[29] Sorkine O, Cohen-Or D. Least-squares meshes. In: Proceedings of shape modeling international, IEEE Computer Society Press; 2004. p. 191–9.
[30] Martn-Merino M, Munoz A. A new sammon algorithm for sparse data visualization. In: Proceedings of the 17th international conference on pattern recognition (ICPR'04); 2004.
[31] Fekete J-D, Grinstein G, Plaisant C. IEEE InfoVis 2004 contest, the history of InfoVis. ⟨http://www.cs.umd.edu/hcil/iv04contest⟩; 2004.
[32] Aamodt A, Plaza E. Case-based reasoning: foundational issues, methodological variations, and system approaches. AI Communications 1994;7(1):39–59.
[33] Agrawal R, Srikant R. Fast algorithms for mining association rules. In: Proceedings of the international conference on very large data bases, VLDB; 1994. p. 487–99.