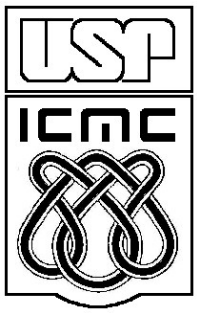
UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação
ISSN 0103-2569

RULE-BASED TOPIC MINING TO ASSIST USER-CENTERED VISUAL EXPLORATION OF DOCUMENT COLLECTIONS

ROBERTO PINHO
MARIA CRISTINA F. OLIVEIRA
ROSANE MINGHIM
ALNEU DE ANDRADE LOPES
RENATO RODRIGUES

Nº 345

RELATÓRIOS TÉCNICOS



São Carlos – SP
Jun./09

Abstract

We propose a three step iterative and interactive visual text mining process to assist users in exploring document collections. In the proposed approach (i) topics are automatically extracted from a document collection , (ii) users explore a similarity-based document map and its related topics, while refining a topic list, and (iii) map quality itself and topic list definition can both be improved based on user interaction. A selective and sequential covering association rule induction strategy is employed to extract the topics. In this strategy, association rules are sequentially induced from selected (manually or automatically) groupings in the similarity-based document maps. Resulting topics are displayed on a Topic Tree control window that assists users in exploring the collection by (i) identifying documents related to specific topics in the map, (ii) removing uninteresting documents from the map, based on their topics, (iii) comparing related topics and documents, (iv) extracting new topics from user selected map regions or from the entire map, (v) building derived maps, and, (vi) eventually exporting sets of labeled documents. Derived maps inherit the previous topic definitions, while benefiting from the removal of undesired documents and, optionally, from the use of terms descriptive of relevant topics to compute document similarity. We present two case studies – on an online news corpus and on a collection of scientific papers – to illustrate our process and its suitability to explore document collections.

1 Introduction

Visual Text Mining (VTM) refers to the analysis of unstructured textual data aided by interactive visual representations capable of inserting user expertise into the mining process. This overall synergistic approach is also a goal for Visual Analytics, which advocates the coupling of visual metaphors and interaction with analytical approaches to assist detection of valuable and possibly unexpected information from dynamic and vast data sources [WT04].

We have been employing association rule (AR) mining in a VTM context for automatic topic extraction from text collections, under the hypothesis that co-occurrence of relevant terms carry semantic content that allows identifying relevant topics. Rule extraction is typically applied to a bag-of-words (vector) representation [SWY75] of the collection. Conventional rule extraction algorithms produce poor results in a topic extraction scenario, due to problems typical of AR mining, such as combinatorial rule explosion and identification of rare associations. Rule redundancy is also a major concern, as sometimes a large number of slightly differing rules refer to the same topic. When handling text one needs a customized solution to deal with these issues.

The proposed approach integrates topic extraction with visual maps of document collections. In particular, we deal with document maps created by projecting the high dimensional vector space model down to two dimensions, while seeking to preserve the documents' original distance (dissimilarity) relationships [Wis99, BCB03, POM07]. In exploring such a map, selecting a group of

clustered points likely corresponds to selecting a set of documents with similar content that address common or related topics. Annotating this map with topics addressed by its documents is important to aid user exploration in applications that require analysis of large document collections.

Our topic extraction approach benefits from this visual grouping by similarity to identify relevant topics from selected regions of the map. A selective association rule induction algorithm, described in Section 3, is steered to only generate rules that include at least one relevant term. Terms are deemed relevant if highly frequent in the selection, as compared to their frequency in the entire document map. Moreover, the topics help users to explore and refine the map, by: (i) identifying regions in the map related to a certain topic, (ii) removing documents on themes deemed not relevant for the user task, (iii) comparing or merging topics, (iv) extracting new topics descriptive of user selected map regions, (v) building new maps and (vi) exporting sets of documents classified by topics during the process once undesired topics/documents have been removed or important topics have been identified.

In this paper, Section 2 provides background information and describes related work on topic extraction and visual text mining. Section 3 details our topic extraction strategy. Navigation and interaction over rules is provided by the Topic Tree control, detailed on Section 4. A discussion follows, in Section 5, of case studies carried out to illustrate the iterative process. Section 6 lists our conclusions.

2 Background and Related Work

2.1 Association Rules in a Text Mining Context

An association rule is an implication of the form $X \Rightarrow Y$, where $X \cap Y = \emptyset$, and both X and Y are subsets of a set of items $L = l_1, l_2, \ldots, l_m$. When mining association rules from text, items are terms from the bag-of-words model, and a transaction T is a set of terms $T \subseteq L$ that represents a document. The rule $X \Rightarrow Y$ holds in a document set C with confidence c if $c\%$ of the documents that contain X also contain Y. The rule $X \Rightarrow Y$ has support $s\%$ in C if $s\%$ of the documents contain $X \cup Y$. In addition to support and confidence, several other objective measures have been defined to evaluate associations in text mining tasks, such as *Interest*, *Conviction*, *Dependency*, *Novelty* and *Satisfaction* [GH06].

Combinatorial rule explosion and finding rare associations [Wei04] are major problems in association rule induction. Broadly, one may handle them: (i) constraining the space of rules to be induced, which is typically achieved by setting high minimum support and confidence; or (ii) filtering the induced rules, based on objective or subjective relevance measures. In either case, it is not uncommon for the process to still output too many rules and, at the same time, miss very interesting ones. If the goal is to identify topics in texts, these problems are particularly critical. A very low minimum support must be set to

find rules that involve both frequent and rare terms, incurring in combinatorial explosion and high rule redundancy. On the other hand, a high support causes potentially interesting rare associations to be completely ignored.

Amir et al. [AAFF05] describe a rule-based approach to search relevant term co-occurrences in text, though not specifically for topic extraction. Closely related terms often appear together in text, and a high number of such co-occurrences can hide relevant, though less frequent, associations of one of such terms with third ones. They handle this with *maximal associations*. Intuitively, a maximal association $X \overset{max}{\rightarrow} Y$ says that whenever X is the only item of its category in a transaction (e.g., *linux*), then Y (e.g. *open-source*) also appears, with some confidence. Identifying such associations with a conventional AR algorithm would require a low confidence threshold, at the cost of also inducing many uninteresting rules. Obtaining maximum association rules is, in general, faster than computing regular ones, nonetheless, it requires a previous step of defining categories of closely related terms.

Cherfi et al. [CNT04], on the other hand, combine several specific measures for post-filtering association rules in a general text mining scenario. They employ several natural language processing tools to automatically extract key-terms from document titles and abstracts. A domain analyst identifies the relevant terms in a manual filtering step. A conventional association rule induction algorithm is then applied and rules are ranked based on multiple metrics defined to reflect different user priorities. The ranking and direct user intervention define which rules will be filtered out.

Topics are closely related to the importance of certain terms in a text collection, and many strategies other than rule induction have been employed for automatic topic extraction. A common approach in text mining is to cluster documents based on similarity, and then identify phrases or keywords representative of each cluster [CKPT92]. The problem also shares many issues with that of keyword identification for document indexing and summarization.

2.2 Topic Extraction in a Visual Text Mining Context

When exploring a map that reflects document content similarity, users need to know the important topics and sub-topics addressed by different groups of documents. In a visual text mining scenario, automatic topic extraction can support: (i) creating maps at multiple abstraction levels; (ii) filtering and focusing through dynamic topic and theme refinement; (iii) enhancing maps with informative annotation; and (iv) showing thematic changes in time varying text collections.

The problem of labeling visualizations with informative topics or terms has been addressed in several text visualization systems. The ThemeView visualization [Wis99] presents a text collection as a landscape view where hills are formed by adding a weighted contribution from document terms within a region. Labels are assigned to hill tops based on those terms that contribute the most to hill height. On a quite different approach, Chen [Che04] employs *Principal Component Analysis* (PCA) to extract the most frequent terms from the

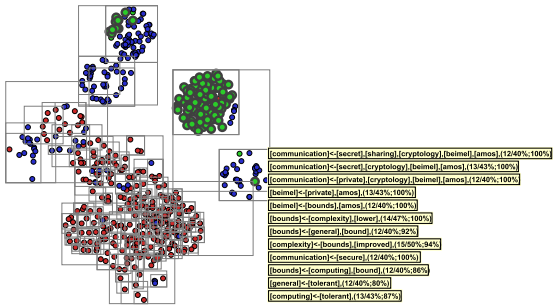


Figure 1: Document map of a corpus of 574 scientific articles. Red, green, and blue glyphs represent papers on Case-Based Reasoning (CBR), Inductive Logic Programming (ILP) and Information Retrieval (IR) respectively. Grey boxes mark automatically selected areas used in topic extraction.

top ten documents related to the first three principal components. The most frequent words are deemed topics and used as captions. Documents are colored and labeled according to their relation to these principal components.

Skupin [Sku02] computes and ranks term frequencies to label regions in document maps that resemble cartographical maps. His maps display the document collections at multiple abstraction levels by means of a hierarchical clustering of the documents. He displays major topics and sub-topics by assigning labels to the different hierarchical levels, based on variations of term frequency evaluation formulas computed for the clusters.

The ThemeRiver visualization shows thematic changes over time for a collection of time tagged documents [HHWN02]. Each theme is shown as a colored river whose width reflects its strength in documents along a particular time period – the strategy to determine theme strength is similar to the one adopted in ThemeView [Wis99]. There are also VTM applications designed to assist topic detection in a single document, such as *Topic Islands* [MWBF98]. It applies wavelet transforms to detect and display topical changes within a document.

3 Topic Extraction using Visualization and Locally Weighted Association Rules

In a previous paper [LPPM07], we introduced an algorithm for automatic detection of the major topics addressed by a sub-set of user-selected documents in a document map. The user delimits a region of the map to select a sub-set of documents, and topics (seen as a set of representative term associations) are obtained in real time and displayed over the map, as illustrated in Fig. 1. The approach extends the classical *Apriori* algorithm [AS94] to perform selective rule induction.

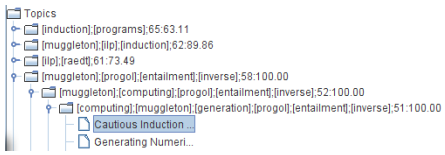


Figure 2: Example of the topic tree control.

The underlying rationale for the proposed algorithm (Algorithm 1) is to restrict the search space to capture the most meaningful rules, without missing relevant rare associations. It starts by identifying relevant terms in the selection, refered to as *seeds*. Seeds are terms with higher frequency in the selection than in the remaining documents in the collection. Then, only rules having at least one seed term are induced. The next step is to weight and rank rules based on their relevance, considering the relative importance of its terms in the selection. In some sense, the approach resembles the rule extraction strategies advocated by Liu et al. [LHM99] and also by Freitas [Fre00], who employ multiple minimum supports to mine association rules in transaction databases.

The weight of a term is given by its relative frequency within the selection, as compared to its frequency in the entire corpus – similarly to the Tf-Idf measure[SB88] and to the ideas described in [BKR98]. Intuitively, terms highly frequent only in the selection are likely to be more informative of its content. In this sense, rule relevance can only be stated comparatively relating the selected documents with the whole corpus. So, similarity-based grouping depicted by the visual map – or actually by any similarity grouping strategy – gains great importance, as seeded rule induction must be performed on groups of similar documents. The weight of a term t_j in a selection S_k taken from a corpus C is given by the summation of its frequencies $Tf_{t_j S_k}$ from documents in S_k divided by the summation of frequencies $Tf_{t_j C}$ from documents of the entire corpus C:

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

To support topic-driven exploration of a corpus, the same rationale of selective rule generation may be applied over the whole document map. Moreover, the earlier approach has been refined to generate an improved rule set for topic extraction, reducing redundancy and supporting topic identification at multiple levels, which is important when a group of documents addresses a global topic and different sub-topics.

The original rule induction algorithm is now embedded into a two-fold sequential covering strategy and extended to handle a whole map, whereas before it was only applied for documents from user selected regions of the map. Rules are still ranked based on the summed local weights of their terms, but rules that cover documents not already covered by better ranked ones are preserved, even if poorly ranked. The resulting rule set captures the most representative topics

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
ranked rules subset size rs
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)
    $W_{t_jS_k} \leftarrow \frac{Tf_{j_{S_k}}}{Tft_{jC}}$ (relative frequency)
    $s(t_j)$ in S_k (support of each 1-itemset)

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

Do
    For each $t_j \in S_k$
        $1\_itemsets \leftarrow 1\_itemsets \cup \{t_j \in Bag\_of\_words | Sup(t_j) > MinSup\}$
        $W_{t_{jS},t_{jC}} \leftarrow \frac{\sum Tf_{t_j,S_k}}{\sum Tft_j,C}$

    %Select n seeds (n terms t_j from 1_itemsets with the larger W_{j_S,j_C})
    $T \leftarrow \{t_1, \ldots, t_n\}$
    $2\_itemsets \leftarrow 2\_itemsets \cup \{(t_i, t_j), i \neq j | (t_i, t_j) \in T \times 1\_itemsets\}$
    $S \leftarrow$Call Apriori for producing rules with $2 <$ number of literals $> m$

    If $S = \emptyset$
        $MinSup \leftarrow 0.75 * MinSup$
        $n \leftarrow n + 1$
While ($S = \emptyset$ and $MinSup >= 0.01$)

$SR \leftarrow \emptyset$
For each $AR_i = (t_{j1} \leftarrow t_{j2}, t_{j3}, \ldots, t_{jm}) \in S$
    $w_{AR_i} \leftarrow \sum_{t_j \in AR_i} \left(w_{t_jS_k}\right)$ % Rule Weight

$Sort(AR, w_{AR})$
$SR \leftarrow \{AR_1, \ldots, AR_{rs}\}$

Return SR

for the group and also additional topics that appear in just a few documents. Simultaneously, rule redundancy is reduced by keeping only those rules that provide additional coverage over previously extracted ones. If two rules cover the same sub-set of documents, only the one with higher summed weight is kept.

The covering strategy (inner loop of Algorithm 2) extracts an initial set of rules, then removes from the selection those documents already covered by them. Rule extraction is repeated on the remaining documents, thus yielding a different rule set, as term seeds are bound to change. This is repeated until no remaining document is left or until no new rules are output. It then adds the set of ranked rules SR extracted from this document sub-set S_k to the Aggregate Rule Set ARS. Function $coverage(AR_i, U)$ on Algorithm 2 computes the number of documents from U that support rule AR_i.

Scanning the whole document corpus requires a map partitioned into groups of similar documents, which may be achieved in two ways. One alternative is to fit a grid over the map and extract rules on each grid cell. The second strategy is to apply clustering the proximity of documents on the map: areas corresponding to different clusters provide the input regions for rule induction. Both strategies introduce an undesired variance, as different grid choices or initial cluster sets will outcome different rule sets. This problem is handled by repeating the rule induction process on slightly perturbed versions of the initial grid, or on different clustering results. This process, referred to as multiple restart, corresponds to the outer loop in Algorithm 2. New rules obtained in each iteration are added to the set of previously extracted ones, until the additional coverage provided by the current iteration is lower than a given minimum or no new rule is added – this we call *multiple restart stop condition*.

Grid based partitioning starts by subdividing the map into r rows and c columns. Each iteration in Algorithm 2 can either increment r and c (usually by 1) or shift the grid. Shifting the grid takes as parameter the number of grid positions s. The initial position is then displaced by $s/cellsize$ in both axes – in this case one extra row and column are added to ensure that the partitioning covers the whole corpus. In this particular case, iterations stop if the total number of grid shifts is reached before satisfying the *multiple restart stop condition*, since these grids would simply overlap previous ones. The k-means algorithm [WF99] is employed for the clustering-based map partitioning – it requires an initial number of clusters and the increment for every iteration in Algorithm 2. In both approaches a maximum number of cells or clusters is also set.

4 Topic-Based Map Exploration

Projection Explorer (PEx) [POM07] is a visual mining platform that incorporates several multidimensional projection techniques and text mining tools to generate document maps based on content similarity. The tool is freely available under a GNU GPL license. We modified it to suit our needs. Fig. 1 depicts a document map generated with PEx.

Algorithm 2 - Sequential covering with multiple selection restart

```
Input: Corpus C                                   % n points,
                                                  % (document × term matrix (M_{n,l}))
Bag_of_words                                      % l literals of matrix
                                                  % document × term
Output: Aggregate Rule Set ARS                    % Set of rules for Corpus C
Algorithm
ARS ← ∅
Define partitioning strategy                      % fast clustering or gridding
Repeat                                            % Outer loop: Multiple Restart
    Partition Corpus C: S ← S_{k_1}, ..., S_{k_n}
    For each cell or cluster S_k
        RuleSet ← ∅
        U ← S_k                                   % uncovered documents
        do                                        % Inner loop: Sequential covering strategy
            SR ← Call Algorithm 1
            NR ← |SR|                             % cardinality of SR
            //rule selection
            while(SR ≠ ∅)
                remove most ranked rule AR_i from SR
                if coverage(AR_i, U) > 0
                    RuleSet ← RuleSet ∪ {AR_i}
                    U ← U− Set of documents covered by AR_i
                end if
             end while
        while(U ≠ ∅ and NR > 0)
        ARS ← ARS ∪ RuleSet
    Refine partition parameters                   % Add cluster or perturb grid
Until coverage(ARS) ≤ previous coverage + δ       % δ usually 1
    Or further partition is not possible or allowed
return ARS
```

To assist topic-based map exploration, extracted rules are displayed in a tree as topics (Fig. 2), each one comprising a set of representative term associations, which describes it, and a set of associated documents, which supports it. Two alternative tree hierarchies are possible. One hierarchy reflects generality by coverage, i.e., a topic B associated to the set of documents D_B, appears as a child of topic A if $D_B \subseteq D_A$. The alternative hierarchy organization reflects the generality relation between topics by their generating term sets, i.e., a topic B described by the set of terms T_B is as a child of topic A if $T_B \supseteq T_A$, being T_B more specific than T_A. In either case, topics covering most documents, typically more general and with fewer terms, appear closer to the root, whereas more specific topics are placed further down. Documents that address a particular topic are displayed as leaf children of the topics they support.

Users may sort nodes either by support or alphabetically (by its first term). Sorting is performed at each tree level. They can also search topics by keyword and export the topic tree as an XML file. Users may select a topic or group of topics in the tree, causing their supporting documents to be highlighted on the document map, which allows them to identify interesting areas on the map. Double clicking a topic opens a dialog box showing the contents of all its supporting documents. Documents supporting selected topics can also be exported to be further analyzed manually or by other tools.

Users may also choose to delete some topics and/or their related documents from the map. Once undesired topics/documents are removed, a new map may be built from the remaining documents and their updated bag-of-words model. Alternatively, a map may be built from a reduced bag-of-words formed only by terms found on topics deemed relevant, which may result in improved maps for exploration, as we have done in the case studies (Section 5).

Two similarity matrix visualizations support user-driven comparison and analysis of the topics in the topic tree. In the *topic similarity view* (Fig. 4) each matrix cell given by $[row(A), column(B)]$ depicts the overlap between documents from both topics A and B – cell color reflects a *Jacquard coefficient* computed over the sets of documents D_A and D_B that support topics A and B:

$$jcoef(A, B) = \frac{|D_A \cap D_B|}{|D_A \cup D_B|}$$

This view conveys the document overlap by different topics.

The *document similarity view* (Fig. 5) compares sets of documents that support different topics. the difference from the *topic similarity view* is that in a cell $[row(A), column(B)]$, A and B being topics, instead of having just one value ($jcoef(A, B)$), we have an embedded matrix comparing documents by topic A, as rows, with documents covered by topic B, as columns. Thus, a cell of one of these embedded matrices reflects the content similarity between two documents. In this view, documents may appear more than once, i.e., in more than one row and column, if they support multiple topics. Now, cell color reflects the content similarity of both documents, as evaluated by the same dissimilarity measure employed to compute the map's underlying projection.

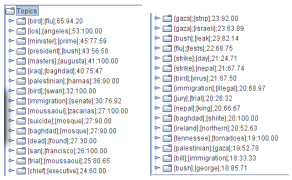


Figure 3: Partial list of topics extracted for the RSS news corpus, sorted by support, top elements shown.

Though they may appear visually similar, these representations convey complementary information. Diagonal cells are always black in the topic similarity view ($jcoef(A, A) = 1$), as they refer to identical supporting document sets. In the *document similarity view*, comparison is pairwise by documents, and a black diagonal is expected to be found as it compares each document with itself. However, the embedded matrix that compares documents from a topic with themselves results in a darker area if the documents that address the topic are highly similar in content, and in a lighter area otherwise. Empirical results show that topics that correspond to broad, poorly informative term sets, are often displayed as very light areas. Also, one may eventually find two topics that share just a few documents in common (or none at all), yielding a very low Jacquard coefficient, yet their supporting documents may be very similar. Such cases would be represented as a white or very light cell in the topic view, while the document view would show a few darker areas or bands where sets of similar documents from each topic cross each other.

5 Case Studies

We describe two case studies that illustrate how our solution supports user-centered analysis of document collections. The first one was conducted on a corpus of 2,684 RSS news feed articles, collected from BBC, Reuters, CNN, and Associated Press sites, during two days in April 2006, henceforth the RSS news corpus. Stop words were removed and frequency-based Luhn's cuts applied to select relevant terms prior to further processing.

The user task was to identify the three most active stories for the period from this very diverse corpus. The topic-based user-driven map exploration comprised the following steps:

1. *Map construction*, using the cosine distance and the ProjClus projection technique to generate the map.

2. *Topic induction.* Topics were induced for the entire map using the algorithm described in Section 3 with the clustering strategy for multiple restart (see partial results in Fig. 3). The initial and maximum number of clusters was set to 17 and 189, respectively, with an increment of 86.

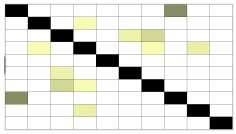


Figure 4: Topic similarity views for topics [bird, flu] through [moussaoui, zacarias] from Fig. 3. Darker colors depicts higher coefficients.

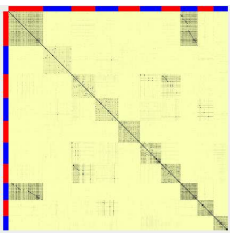


Figure 5: Document similarity view for topics [bird, flu] through [moussaoui, zacarias] from Fig. 3.Darker colors depicts higher similarity.

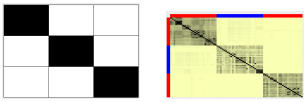
These settings were chosen based on previous ad-hoc experiments, where we found that values around $\sqrt{sizeofcorpus}$ usually yield good results. The process stopped after 3 runs, before reaching the *multiple restart stop condition*. A total of 522 topics were output and displayed in the topic tree. 1317 documents were not covered by any rule – the minimum support for rule induction was set to 3 documents;

3. *User-driven topic analysis*. This task has been sub-divided in two: (i) removal of irrelevant topics (and in some cases also of their corresponding documents); (ii) topic generalization by merging related topics.

4. Rebuilding the map. Documents covered by selected topics are used to generate a new map.

The User-driven topic analysis was performed as follows: iteratively, we inspected the top 10 or 20 topics, sorted by support, considering their similarity matrices views and checking the distribution of their corresponding documents over the map – documents scattered throughout the map might signal a poor, non descriptive, topic. After each inspection, topics could be merged, if similar, or simply removed, if deemed not significant. Eventually, document titles where inspected.

Figures 4 and 5 show similarity matrices for the top 10 topics initially extracted (Fig. 3). From the topic view, one observes that topics [bird, flu] (1^{st}) and [bird, swan] (8^{th}) are closely related, which is confirmed in the document view by the diagonal line of black pixels found when comparing both topics.



(a) (b)

Figure 6: Topic and Document similarity views for topics [baghdad, mosque], [moussaoui, zacarias] and [dead, swan], respectively.

This line is the result of comparing documents with themselves, when documents are covered by both topics. From the document similarity view(Fig. 5), one notices also that documents that support topic [los, angeles] (2^{nd}) are not as similar to one another as those that support [bird, flu], as the area given by comparing documents from topic [los, angeles] is lighter then the area given by the comparison of documents from [bird, flu] with themselves. Further inspection showed that "Los Angeles" is merely the place of many disconnected events, so the topic was removed.

The topics [baghdad, mosque] and [moussaoui zacarias] cover no (or just a few) documents in common (Fig. 6(a)), but still are somewhat related as their respective documents are similar. Notice in Fig. 6(b) how the document sets from these two topics (cells $(1, 2)$ or $(2, 1)$) have darker areas than comparing these two topics with [dead, swan] (cells $(3, 1), (3, 2), (1, 3), (2, 3)$). After a few topic merges and removals, the top three topics narrowed down to [bird, flu, swan, dead, found,tests], [palestinian, hamas, gaza, strip, israeli], and [iraq, baghdad, suicide, mosque]. To further include documents on these topics, we searched topics that shared at least one term with them and compared results using the matrices. Again, if similar, they were merged, resulting in the following set of topics: [bird, flu, swan, dead, found, tests, virus], [palestinian, hamas, gaza, strip, israeli, shot, murdered, soldier, cameraman, jury, british], and [iraq, baghdad, suicide, mosque, shiite, affiliated, bomb, bombers].

Finally a new map was produced, only with documents related to the selected three top topics. This new map placed documents from each topic in different regions. A new set of topics was extracted, which revealed important details related to each main topic. For instance, the topic [mosque,bombers,women] related to the general topic about a suicide bombing at a mosque in Iraq, refers to news covering the fact that bombers from that suicide attack were disguised as women.

The second study was performed on a collection 574 scientific articles (titles, abstracts and references only) manually classified into one of three subject areas: Case-Based Reasoning (CBR), Inductive Logic Programming (ILP) and Information Retrieval (IR), henceforth the CBR-ILP-IR corpus. A superset of this corpus has been employed to compare multidimensional projection techniques [PNML08]. It was processed in the same fashion as the RSS news corpus.

As it can be noticed on Fig. 1, the map does a good job on visually grouping documents from each class. However, manual classification was a demanding

task, one that without the support of automated tools demanded reading each document and assigning a class.

On this second case, we show that applying our topic driven exploration strategy can give a user an understanding of the corpus, while simultaneously producing a topic definition that closely matches classes that were manually produced, without the burden of examining every single document. For this task, we did not color glyphs by class, to resemble a scenario where such classification is not available.

A user familiar with the subject areas covered by the corpus explored it, following the same task outline defined for the previous case study, although he was free to focus on topics from the tree that caught his attention. Following evidence from the topic similarity view, he realized that topics [muggleton, raedt], [muggleton, ilp], [muggleton, entailment], and [muggleton, order] are covering quite similar sets of documents. Thus, they were merged into [muggleton, ilp], as he decided to label the new topic. Comparison of this new aggregate topic with topics [ilp, raedt], [induction, raedt] and [raedt, dzeroski], lead to a new aggregate topic, which he named [inductive logic programming], knowing that Raedt, Muggleton and Dzeroski are important authors from the field and after he inspected some of the articles. This higher level topic was then compared with the previous manual classification, with very satisfactory results: The topic is supported by 116 documents, 114 of those previously labeled as ILP. The ILP class has a total of 119 documents. In Fig. 1, documents covered by the aggregate [inductive logic programming] topic are shown with thicker borders, closely matching the set of manually labeled ILP documents (in green).

6 Conclusions

We introduced a framework, added to the PEx open-source platform, to support user-assisted exploration of large document collections. Two elements, topic extraction over the document map and topic-supported map exploration and refinement, combine into a valuable tool for users to gain understanding and extract relevant information from complex document collections.

The whole process is based on an association rule induction algorithm devised to effectively extract topics – described as term associations – from text. Because it scans the space of documents based on their content similarity and adopts a suitable criterion to identify relevant terms within groups of similar documents, the algorithm manages to find both frequent and rare term associations that appropriately describe topics. Moreover, it outputs a rule set that is representative of the topics addressed in the whole corpus. The coverage of the whole corpus is achieved by applying a multiple restart partitioning strategy that handles the problem of variance that could arise when adopting a sequential coverage strategy for induction.

Functionality is provided to enable users to explore the corpus based both on the document map and on the extracted topics. Topics are displayed in a topic tree and may be deleted, edited, merged and compared, as users browse

them and inspect related documents. Users may also purge documents from the collection based on their topics, and generate new maps that reflect their previous interventions on both the corpus and the topic list. Maps that, e.g., include only documents related to remaining relevant topics may then undergo a new round of topic extraction and exploration, in an iterative and interactive visual mining process that feeds on itself and on user input to enhance analysis capability.

References

[AAFF05] Amir A., Aumann Y., Feldman R., Fresko M.: Maximal association rules: A tool for mining associations in text. *J. Intell. Inf. Syst. 25*, 3 (2005), 333–345.

[AS94] Agrawal R., Srikant R.: Fast Algorithms for Mining Association Rules in Large Databases. *Proceedings of the 20th International Conference on Very Large Data Bases* (1994), 487–499.

[BCB03] Börner K., Chen C., Boyack K.: Visualizing Knowledge Domains. *Annual Review of Information Science and Technology(ARIST) 37* (2003), 179–255.

[BKR98] Bookstein A., Klein S., Raita T.: Clumping properties of content-bearing words. *Journal of the American Society for Inf. Science 49*, 2 (1998), 102–114.

[Che04] Chen C.: *Information Visualization: Beyond the Horizon*. Springer-Verlag London Ltd, 2004.

[CKPT92] Cutting D., Karger D., Pedersen J., Tukey J.: Scatter/Gather: a cluster-based approach to browsing large document collections. *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval* (1992), 318–329.

[CNT04] Cherfi H., Napoli A., Toussaint Y.: Towards a text mining methodology using frequent itemsets and association rules. *Soft Computing Journal 11* (2004).

[Fre00] Freitas A.: Understanding the crucial differences between classification and discovery of association rules: a position paper. *ACM SIGKDD Explorations Newsletter 2*, 1 (2000), 65–69.

[GH06] Geng L., Hamilton H. J.: Interestingness measures for data mining: A survey. *ACM Comput. Surv. 38*, 3 (2006), 9.

[HHWN02] Havre S., Hetzler E., Whitney P., Nowell L.: Themeriver: Visualizing thematic changes in large document collections. *IEEE Transactions on Visualization and Computer Graphics 8*, 1 (2002), 9–20.

[LHM99] Liu B., Hsu W., Ma Y.: Mining association rules with multiple minimum supports. In *KDD '99: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining* (New York, NY, USA, 1999), ACM Press, pp. 337–341.

[LPPM07] Lopes A. A., Pinho R., Paulovich F. V., Minghim R.: Visual text mining using association rules. *Comput. Graph. 31*, 3 (2007), 316–326.

[MWBF98] Miller N., Wong P., Brewster M., Foote H.: TOPIC ISLANDS TM-a wavelet-based text visualizationsystem. *Visualization'98. Proceedings* (1998), 189–196.

[PNML08] Paulovich F. V., Nonato L. G., Minghim R., Levkowitz H.: Least square projection: A fast high-precision multidimensional projection technique and its application to document mapping. *IEEE Trans. on Visualization and Comp. Graphics 14*, 3 (2008), 564–575.

[POM07] Paulovich F. V., Oliveira M. C. F., Minghim R.: The projection explorer: A flexible tool for projection-based multidimensional visualization. In *Proceedings of the XX Brazilian Symposium on Computer Graphics and Image Processing - SIBGRAPI* (Belo Horizonte, Brazil, 2007), IEEE CS Press, pp. 27–36.

[SB88] Salton G., Buckley C.: Term-weighting approaches in automatic text retrieval. *Information Processing and Management: an International Journal 24*, 5 (1988), 513–523.

[Sku02] Skupin A.: A cartographic approach to visualizing conference abstracts. *Computer Graphics and Applications, IEEE 22*, 1 (2002), 50–58.

[SWY75] Salton G., Wong A., Yang C. S.: A vector space model for automatic indexing. *Commun. ACM 18*, 11 (1975), 613–620.

[Wei04] Weiss G. M.: Mining with rarity: a unifying framework. *SIGKDD Explor. Newsl. 6*, 1 (2004), 7–19.

[WF99] Witten I., Frank E.: *Data Mining:: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 1999.

[Wis99] Wise J.: The ecological approach to text visualization. *Journal of the American Society for Information Science 50*, 13 (1999), 1224–1233.

[WT04] Wong P., Thomas J.: Visual Analytics. *Computer Graphics and Applications, IEEE 24*, 5 (2004), 20–21.