International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:1, No:11, 2007

Bottom Up Text Mining through Hierarchical Document Representation

Y. Djouadi., F. Souam.

***Abstract*—**Most of the existing text mining approaches are proposed, keeping in mind, transaction databases model. Thus, the mined dataset is structured using just one concept: the "transaction", whereas the whole dataset is modeled using the "set" abstract type. In such cases, the structure of the whole dataset and the relationships among the transactions themselves are not modeled and consequently, not considered in the mining process.

We believe that taking into account structure properties of hierarchically structured information (e.g. textual document, etc …) in the mining process, can leads to best results. For this purpose, an hierarchical associations rule mining approach for textual documents is proposed in this paper and the classical set-oriented mining approach is reconsidered profits to a Direct Acyclic Graph (DAG) oriented approach. Natural languages processing techniques are used in order to obtain the DAG structure. Based on this graph model, an hierarchical bottom up algorithm is proposed. The main idea is that each node is mined with its parent node.

***Keywords*—**Graph based association rules mining, Hierarchical document structure, Text mining.

I. Introduction

A large portion of the commonly available information does not appear in structured databases but in textual form and text still remains an essential way for conveying information. For this purpose, text mining has emerged as a new and promising research area of text processing and data mining. Text mining is defined [6] as: "knowledge discovery in textual data as articles, newswires, reports, Web pages, etc…" and is focused on the discovery of new facts and world knowledge from texts that do not explicitly contain the knowledge to be discovered.

The goals of text mining are some similar to those of data mining: for instance, it also attempts to find cluster, uncover trends, discover associations, etc … Association rules mining are usually used as text mining techniques. These discovered associations are of general purpose. In [8] these associations between terms are used to expand queries trough the construction of a knowledge base. In [11], the proposed mining process founds relevant associations and computes the relevance of the result on the following way :

query : "*find all associations between a set of countries including Iran and any person*"
results : [*Iran, Nicaragua, USA*] →*Reagan* (6/1.000)

There are several data mining researches to discover interesting rules or patterns from well-structured data such as transaction databases with boolean or numeric attributes. However, it is difficult to directly apply the traditional data mining techniques to a text since this text consists of (i) heterogeneous data and (ii) unstructured or semi-structured data. Therefore, there still are a few number of studies on text mining [4]. Hence, the necessity for specialized techniques specifically operating on unstructured or semi-structured textual data.

Recently in all text-oriented applications, including text mining, there is a tendency to start using more complete representations than just nouns or keywords, i.e. representations with more types of textual elements. In text mining, for instance, there is the belief [9] that these new representations will expand the kinds and enhance the relevance of discovered knowledge.

We believe also that hierarchical organization is an natural and expressive representation of textual information. Indeed, an important part of the encountered textual information (usenet, scientific papers, laws journal, books, etc …) follows this kind of representation. Moreover of the syntactical representation, implicit knowledge (e.g. inclusion relationship, specialization/generalization relationship etc…) is conveyed through the choice of an hierarchical organization. Such knowledge can be qualified as semantic. Even though, a pragmatic aspect may be attached to hierarchical organization, essentially for a textual document.

For this purpose, we propose in this paper, an N.L.P. (Natural Language Processing) based approach which extract the document structure in the first phase, underlying that a text mining process consists of two phases. A direct acyclic graph (DAG) model of document is then formalized. Based on this graph model, an hierarchical bottom up mining algorithm is proposed in order to discover associations rules in a given text (second phase).

The rest of this paper is organized as follows. Section 2 presents some developments of text mining and propose our general framework for mining hierarchical textual documents. Section 3 presents syntactical concepts for extracting document structure. It also proposes a semantic for text structure.

A hierarchical model of textual document is proposed in section 4. Based on this model, the graph oriented method is presented in the next section. The last section gives conclusion

Y. Djouadi. and F. Souam are with the Department of Computer Science. State University of Tizi Ouzou, Algeria (e-mail: ydjouadi@mail.ummto.dz).

and orientations for future works.

II. General Framework

Text mining [9] consists of two main phases: a first pre-processing phase and a mining phase. In the first phase, all portions of text are transformed to some kind of intermediate representation that allows their automatic analysis. In the second phase, the intermediate representations are analyzed and interesting and non-trivial patterns are discovered.

A. Pre-Processing Text: General Trends

Semantic pattern, concepts hierarchy, intermediate structure and so one techniques have been already used as knowledge based text pre-processing approaches.

In [7] an approach which utilizes concepts and their hierarchy is proposed. This addresses the problem of forcing a structure to unstructured text documents so the data mining techniques can be applied to them. The concept hierarchy is structured as a directed graph of concept relationships. For example, the relationship "*communication devices* → *phone*" shows that *communication devices* is a more general concept than phone. The hierarchy can be customized such that it only contains concepts and relationships that may be of relevance to the user. The concept hierarchy is then used to tag words or phrases within the text, tagging it with the relevant concept and implicitly tagging all of the ancestors of that concept in the process. Here, authors do not fix any method of tagging. Pointing out that many text processing methods can be used. Sets of concepts will be generated for each document and relationships within and between are discovered using statistical approaches. On the other hand, in this approach, a domain specific structure is needed. This structure is hand crafted.

Intermediate representations are also used in [10]. Authors propose a method that uses conceptual graphs (more exactly Sowa graphs). This method consists of: (i) the transformation of texts into conceptual graphs and (ii) clustering the set of conceptual graphs into several groups and organizing them into a hierarchy.

In [12], the proposed approach consists of representing the grammatical structure of text sentences for information extraction. Authors hypothesize that incorporating the sentence structure instead of representing the sentence as a sequence of tokens, results in better extraction accuracy.

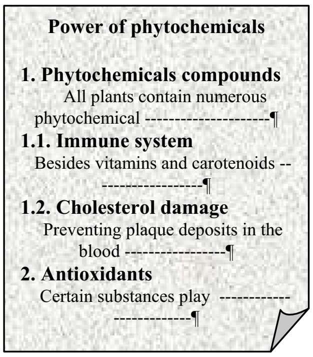
Beyond text mining, structure based pattern mining is also addressed. In [1] the mined structure consists of a constrained graph. This graph models newsgroups (e.g. alt.business.insurance) where the posting and its respective response consist of connected vertices. One the most relevant approach is proposed in [15]. A taxonomy (*is-a* relationships) is considered over the items. Transactions are expanded in order to integrate ancestors in the taxonomy of each item. Authors have developed an interest measure based on this taxonomy which permits to prune 40% to 60% of redundant rules. The interest of this approach is clearly illustrated. However we think that pre-processing the input dataset in order to extract such taxonomy is a harder process.

Concluding remarks:

Concepts hierarchy (taxonomies) is a strong way for text mining. However, inducing implicit taxonomies may be quite difficult for two main reasons:

1) Obtaining such representation leads to another open problem (e.g. semantic networks , grammatical structure, etc …).
2) Multiple taxonomies may be present for a same domain.

Hence it may be remarked that most textual documents are already hierarchically organized and represented through simple syntactical marks (e.g. headings, sub-headings, etc …). The goal of our proposed approach is to use this already existing structure as a significant representation without need to harder pre-processing text mining phase as it is usually proposed by the community. It may also be remarked that the document structure may easily be assimilated to a concept hierarchy.



Power of phytochemicals

1. Phytochemicals compounds
All plants contain numerous phytochemical ---------------------¶
1.1. Immune system
Besides vitamins and carotenoids -------------------¶
1.2. Cholesterol damage
Preventing plaque deposits in the blood -----------------¶
2. Antioxidants
Certain substances play ------------------------¶

Fig. 1 Textual document

Thus, the graph-model based approach proposed in this paper consists of two main phases:

1) Document structure extracting phase.
2) Associations rules discovering phase.

B. Document Structure Extracting

N.L.P. (Natural Languages Processing) techniques are used in our approach in order to preprocess the mined document. Whereas textual units marks (see §III.B) gives the structure of document. Complete details of the method are given in [13] (objective for this paper is the mining process).

The structure extracting process leads to a direct acyclic graph structure (see fig. 1 and 2). The nodes of the structure consists of textual units which are logical entities (chapters, introduction, section, part, definition, etc…). These logical entities are considered regarding their intrinsic structure and layout properties.

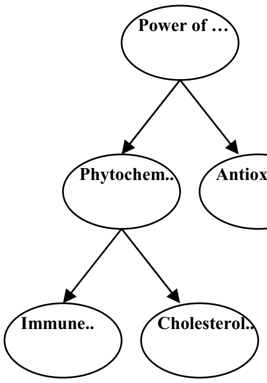


Fig. 2 Document structure model

C. Associations Rules Discovering

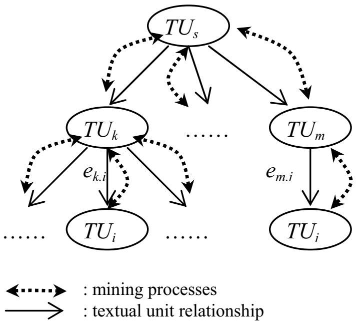
The mining process is based on the document structure model. An oriented edge $e_{k,i}$ (see fig. 3) indicates a relationship $R_v(k,i)$ between its adjacent vertices (i.e. the textual units TU_k and TU_i) where the TUs are obtained from the pre-processing phase. Then, the proposed method gains in terms of semantics and computational complexity.

Semantics: a relationship $R_v(x,y)$ means a semantic link between textual units TU_x and TU_y. Thus, the main idea consists of mining between them those textual units which are strongly semantically linked i.e. those which verify the R_v relationship.

Computational complexity: Let us consider a classical associations rule mining algorithm like APRIORI [2]. Let C_k denote the set of all potentially k-itemset ($C_k = F_{k-1} \bowtie F_{k-1}$). Computing C_k for a given set F_{k-1} is roughly an $\sum_p O(m^2)$ operation, where m is the number of identical itemsets in F_{k-1} and p is the cardinality of F_{k-1}.

In our proposed method, computing C_k becomes roughly an $\sum_t \sum_p O(a^2)$ with $a << m$ (t is the number of textual units).

Fig. 3 Graph oriented mining process



III. Interpretation of Document Structure

Properties of text structure has been investigated essentially by the linguistic community. Theses studies concern essentially natural language processing domain and related objectives such as automatic comprehension. In [5], the marks of text structure are considered for textual unit decomposition.

A. Semantic Aspect

A textual document can be viewed as a particular organization of logical entities. This text structure reflects some semantics since it contributes to the comprehension of the considered document. Also, text structure allows the following to be perceived :

1) Logical entities: chapters, sections, paragraphs, lists, introduction, but also definitions, commentaries, theorems, demonstrations; etc…
2) Relationships between these logical entities: inclusion, semantic link (theorem / demonstration, textual object / its commentary), logical link (navigational link); etc…

In our approach, two precedence relationships between textual units will be considered. The former relationship (denoted $x \prec y$, where x and y are textual units) indicates that if x is a k-order heading then y is a k+1-order heading. The latter relationship (denoted $x \equiv y$) indicates that x and y are same order heading.

B. Syntactical Aspect

It is generally agreed that beyond classical text punctuation (e.g. "**.**", "**;**", "**:**", etc …) there is higher level morphological, lexical and layout marks (indentation, font changes, etc …) which are linking a series of sentences to form particular textual entities (e.g. chapter, section, part) and expressing particular relationships between these entities [14]. The set of textual entities and their relationships defines the structure (organization) of the text.

Sequences of structure marks, in a given order, leads to a certain interpretation in terms of text structure.

It results that characterizing the text structure imposes defining structure marks. In this paper, we will consider marks allowing certain types of textual entities to be distinguished and structured as textual punctuation, that is :

1) body size and/or style of characters,
2) indents and new lines,
3) headings such as *chapter* (i.e. lexical mark),
4) differences and positional marks applied to these headings,
5) typographical marks.

Definition 1. a textual unit mark is well formed sequence of structure marks. An order p (p ≥ 0) is assigned to each textual unit mark.

Definition 2. a textual unit is a textual entity delimited by two textual unit marks.

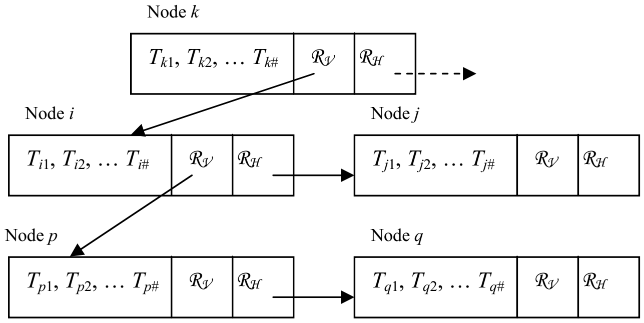

Node k
$T_{k1}, T_{k2}, \dots T_{k\#}$
$\mathcal{R}_\mathcal{V}$
$\mathcal{R}_\mathcal{H}$
Node i
$T_{i1}, T_{i2}, \dots T_{i\#}$
$\mathcal{R}_\mathcal{V}$
$\mathcal{R}_\mathcal{H}$
Node j
$T_{j1}, T_{j2}, \dots T_{j\#}$
$\mathcal{R}_\mathcal{V}$
$\mathcal{R}_\mathcal{H}$
Node p
$T_{p1}, T_{p2}, \dots T_{p\#}$
$\mathcal{R}_\mathcal{V}$
$\mathcal{R}_\mathcal{H}$
Node q
$T_{q1}, T_{q2}, \dots T_{q\#}$
$\mathcal{R}_\mathcal{V}$
$\mathcal{R}_\mathcal{H}$

Fig. 4 Model implementation

IV. Formalizing Document Structure Model

We have seen that particular structure marks leads to an hierarchical interpretation and corresponding relationships. Conceptually, a textual document can then be modeled using a direct acyclic graph structure (i.e. a set of nodes, a root node and a set of oriented edges). More formally, a model of document (denoted $\mathcal{M}$) is given as: $\mathcal{M}(\mathcal{N},\mathcal{R}_\mathcal{H},\mathcal{R}_\mathcal{V})$.

Definition 3. $\mathcal{N}$ is the set of textual units. Each element of $\mathcal{N}$ consists of a pair (d,k) where d is the textual content and k is a meta-knowledge associated to d.

Remark: k is a generic concept which permits to consider meta-knowledge (i.e. titles, key words) in the mining process. We are achieving that in our method by increasing the support of the considered meta-knowledge.

Definition 4. The relation $\mathcal{R}_\mathcal{H}$ (respectively $\mathcal{R}_\mathcal{V}$) is defined in the Cartesian product $\mathcal{N}\times\mathcal{N}$ as follows:
let $(x,y) \in \mathcal{N}\times\mathcal{N}$ with $x \neq y$,
then :

$$x \, \mathcal{R}_\mathcal{H} \, y \Leftrightarrow x \prec y$$

$$x \, \mathcal{R}_\mathcal{V} \, y \Leftrightarrow x \equiv y$$

V. Details and Evaluation

We are describing in this section, models, technical implementation and results of our method. We are also discussing obtained results.

A. Models and Algorithms

The graph structure is technically implemented using the above described model. For instance, the document presented in figure 1 leads to the structure presented in figure 4. Each node comprises the set of corresponding transactions, an R_v link and an R_h link. For the first version of HBU_Miner presented in this paper, meta-knowledge are not considered. The R_v link indicates that the pointed node is the child of the considered node, the R_h link indicates that the node and the pointed node have the same parent node.

The proposed method attempts first to construct the graph structure of the document. Then the graph is examined from the leaves, up to the root. At each step, the node and its parent are mined together. Discovered association rules are added at each step. The Examine_DAG algorithm is a recursive one and is given as follows:

```
Procedure Examine_DAG(node, parent_node)
Begin
if ($\mathcal{R}_\mathcal{V}$_Link(node) != null) then
   Examine_DAG($\mathcal{R}_\mathcal{V}$_Link (node), parent_node);
end if
Mining_Nodes(node, parent_node);
if ($\mathcal{R}_\mathcal{H}$_Link(node) != null) then
   Examine_DAG(H_Link(node), Parent_node);
end if
End
```

This procedure has to be called just once with the root node and an empty node (called Φ node) as parameters like :
Examine_DAG(root_node, Φ);

The *Mining_Nodes* procedure implements the classical Apriori algorithm [2]. Devised on the method presented in [3] an hash tree is also used in order to avoid join operation. At each call, this procedure adds the locally discovered frequent itemsets to a global set of frequents itemsets. These itemsets are constituted by nouns. Concerning the transactions. Sentences and paragraphs were individually experimented as transactions. Our first experiments have shown that the use of sentences as transactions give the best results especially in term of variety of associations. So we handle a transaction as a sentence.

B. Experimental Results

The architecture used in our experiments is a 1.8 GHz Pentium IV PC with 248 MB main memory running Windows

XP. The system is fully developed in Java 2.0.

We have considered different categories of documents in our experiments. Table 1 describes two documents. The results for these documents are illustrated through the run time performance. Figure 5 shows that the hierarchical bottom up mining method gives best performances than a global mining method (a full document mining method). However, for documents with important number of nodes, the gain in term of computational aspect in not clear due to the recursive implementation of the algorithm in java. Concerning the relevance of the results, discovered associations has been presented to some users (members of the laboratory). It has been remarked that more relevance is achieved when:

1) The together mined textual units express different concepts.
2) The sentences are more longer.

VI. Conclusion and Future Works

In this paper we have proposed to reconsider the set oriented approach especially for already semi structured documents. We have outlined the interest of considering the document structure in the mining process. Our performance study shows that a hierarchical bottom up mining process outperforms a flat mining process.

TABLE I
TEXTUAL DOCUMENTS

	Document 1	Document 2
Title	"*The power of phytochemicals*"	"*Prevention of breast cancer*"
Document domain	Biochemical	Medicine
Number of words	1.247	32.599
Number of nodes	9	47

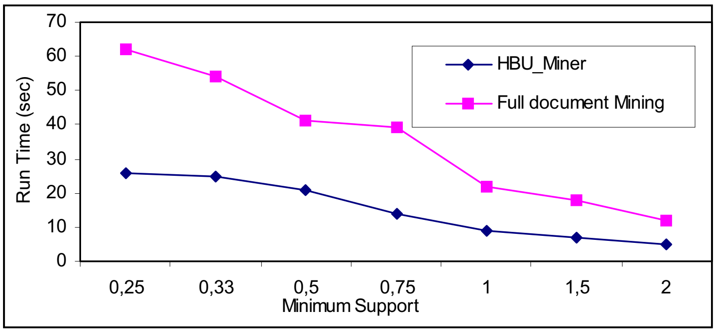


Fig 5.a. Runtime comparison between full document mining and hierarchical mining for document 1.

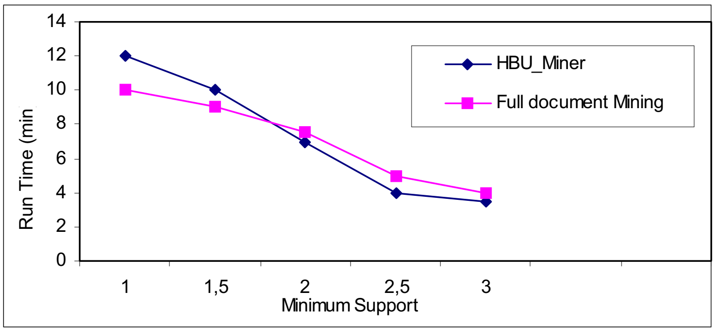


Fig. 5.b. Runtime comparison between full document mining and hierarchical mining for document 2.

Future works will focus on the two following aspects:

1) Including meta-knowledge (titles, etc…) in the mining process.
2) Based on the principle that the document structure model may easily be assimilated to a concept hierarchy. We intend soon to define and compute a semantic distance between textual units. Considering this distance in the mining process, two nodes will be mined together if the semantic distance between them if greater than a threshold to determine. Otherwise the two nodes will be considered as a single node.

References

[1] R. Agrawal, S. Rajagopalan, R. Srikant, Y. Xu, “Mining Newsgroups Using Networks Arising From Social Behavior”, Proceedings of the Twelfth Int’l World Wide Web Conference, Budapest, Hungary, May 2003.

[2] R. Agrawal, T. Imielinski, A. Swami, “Mining associations rules between sets of items in large databases”, In Proc of the ACM SIGMOD Conference on Management of Data, Washington, D.C., 1993, pp. 207-216.

[3] F. Berzal, J.C. Cubero, N. Marin, J.M. Serrano, “TBAR: An efficient Method for Association Rule Mining in Relational Databases”, Data Mining and Knowledge Discovery Journal, Kluwer Academic Publishers, vol 6, 2002.

[4] M. Craven, D. DiPasquo, D. Freitag, A. McCallum, T. Mitchell, K. Nigam, S. Slattery, “Learning to Construct Knowledge Bases from the World Wide Web”, Artificial Intelligence Review, vol 118, 2000, pp 69-114.

[5] R. Dale, “Exploring the Role of Punctuation in the Signalling of Discourse Structure”, Workshop on Text Representation and Domain Modelling, Berlin, 1991 pp. 110-120.

[6] R. Feldman “Mining Text Data”, Chapter 21 in Handbook of Data Mining, Lawrence Erlbaum Associates, 2003, 48 pages.

[7] R. Feldman, H. Hirsh, “Finding Associations in Collections of Text”, Machine Learning, Data mining and Knowledge Discovery: Methods and Application, In R.S. Michalski, I. Bratko, and M. Kubat editor, John Wiley and Sons Ltd, 1997.

[8] M.H. Haddad, J.P. Chevallet, M.F. Bruandet, “Relations between Terms Discovered by Association Rules”, European Conference on principles and Practices of Knowledge Discovery in Databases, PKDD’2000, Lyon, France, September 2000.

[9] Hearst. “Untangling Text Data Mining”. Proceedings of ACL’99, 37th Annual Meeting of the Association for Computational Linguistics, 1999.

[10] M. Montes-Y-Gomez, A. Gelbukh, A. Lopez-Lopez, R. Baeza-Yates, “Text Mining with Conceptual Graphs”, Symposium of Natural Languages Processing and Knowledge Engineering, NLPKE-2001, IEEE, Tucson, USA, October 2001.

[11] M. Rajman, R. Besançon, “Text Mining - Knowledge extraction from unstructured textual data”, Proc. of 6th Conference of International Federation of Classification Societies (IFCS-98), Roma (Italy), July 98, pp 473-480.

[12] S. Ray, M. Craven, “Representing Sentence Structure in Hidden Markov Models for Information Extraction”, Proceedings of the 17th International Joint Conference on Artificial Intelligence. IJCAI 2001.

[13] F. Souam, “Transactions Expansion for Mining Hierarchical textual Documents”. Master Thesis, University of Tizi-Ouzou, Algeria, to appear in 2006.

[14] E. Pascual, J. Virbel, “Semantic and Layout Properties of Text Punctuation”, Workshop on Punctuation in Computational Linguistics, ACL’96, USA, June 1996.

[15] R. Srikant, R. Agrawal, “Mining Generalized Associations Rules”, Proceedings of the 21st VLDB Conference, Zurich, Switzerland, 1995.