Mine Your Own Business, Mine Others' News!

Quang-Khai Pham[†‡], Regis Saint-Paul[†],
Boualem Benatallah[†]
[†]School of Computer Science & Engineering
University of New South Wales
Sydney, NSW 2033, Australia
{qpham, regiss,
boualem}@cse.unsw.edu.au

Noureddine Mouaddib[‡],
Guillaume Raschia[‡]
[‡]Atlas Group
LINA at University of Nantes
Nantes, France
{noureddine.mouaddib,
guillaume.raschia}@univ-nantes.fr

ABSTRACT

Major media companies such as The Financial Times, the Wall Street Journal or Reuters generate huge amounts of textual news data on a daily basis. Mining frequent patterns in this mass of information is critical for knowledge workers such as financial analysts, stock traders or economists. Using existing frequent pattern mining (FPM) algorithms for the analysis of news data is difficult because of the size and lack of structuring of the free text news content. In this article, we demonstrate a comprehensive **S**treaming **TE**mpor**A**l **D**ata ($STEAD$) analysis framework for mining frequent patterns in financial news. In this demonstration, we show how the mining task is supported by the use of a Time-Aware Content Summarization algorithm (TACS). This summary generates a concise representation of large volume of data by taking into account the expert's peculiar interest while preserving the news arrival temporal information which is essential for FPM algorithms. We experimented the whole framework on a set of news data from Reuters.

1. INTRODUCTION

Frequent sequence mining identifies sequences of events (here called patterns) that occur over time for a significant number of objects. For example, a pattern could be "a salary increase" followed by "a house loan subscription" for a number of bank customers. In this example, the observed objects are customers and each customer history forms a distinct time sequence. A pattern is said frequent if it is observed in more than a given ratio (called support) of all customer histories.

The discovery of such pattern is of importance since the fact that several events are frequently observed in sequence may indicate a correlation or dependency among these events. By observing the first few events in a sequence, an expert may be able to anticipate future events. In our example, the bank could advertise interesting home loans to a customer who just got a salary increase.

In this demonstration, we present our framework for mining frequent patterns in financial news and its application to Reuter's news archives. The application of pattern mining to financial news is challenging for the following reasons.

First, the task is highly dependent on the expert's interest. For instance, consider the takeover of Arcelor by Mittal Steel in 2006. A stock trader could be interested in finding patterns affecting companies in similar sectors (e.g. steel, iron or copper industries) while an economist may be interested in finding correlations between national political turmoil (e.g., in the Arcelor-Mittal case, the Indian and French governments) and corresponding market volatility. Objects of interest are companies in the former example and countries in the latter. A first task is therefore to split the global worldwide stream of news into distinct time sequences related to the expert's objects of interest (e.g. Arcelor, Mittal Steel or France, India). As we will see, the objects of interest might be more or less well defined and a given news item may be related to more than one object (e.g. Arcelor and Mittal Steel), creating specific difficulties during the mining itself.

Second, news are poorly structured data. Some real world news sources are described by over 30 attributes. Several of these attributes—and all the attributes that carry the news' content—are either multivalued or plain text. Thereby, news stories are not directly comparable in terms of their contents, which are again dependent on the expert's specific interests.

Third, interesting patterns are sequences of events such that the order between events has some significance, e.g., "salary increase" followed by a "home loan subscription". Suppose that two news A and B both occur frequently. We will observe that most of the patterns formed by A and B ($\langle A, A\rangle$, $\langle A, B\rangle$, $\langle B, B\rangle$, $\langle B, A\rangle$, $\langle A, A, A\rangle$, $\langle A, A, B\rangle$, etc.) may be frequent. Such patterns however have very limited interest since the fact a pattern $\langle A, B\rangle$ is frequent is a mere coincidence and does not reveal any correlation between A and B: the order between these two events has no significance. Such patterns are called *noisy patterns* and complicate the mining task by creating numerous candidate patterns that need individually to be explored, thereby impacting very significantly the overall performance of mining algorithms.

Fourth, the volume of data considered is large. One month of real world news represents 40000 individual news stories. Depending on the expert's interest, the pattern she might be looking for may span weeks, months or years. For instance, Mittal Steel's bid for Arcelor started in January 2006 and the merger was only announced in June 2006. Unlike numerical data, it is impossible to aggregate news stories to compute

EDBT'08, March 25–30, 2008, Nantes, France.

an "average" of the world events over some period.

In order to address the above challenges, we propose a comprehensive **S**treaming **TE**mpor**A**l **D**ata (*STEAD*) analysis framework for mining of frequent patterns in financial news. We make the following contributions:

- We propose a categorization of news into individual time sequences depending on an expert's defined interests.
- In order to make news content comparable, we allow experts to build a categorical description of news corresponding to their interests.
- We propose a Time-Aware Content Summarization algorithm (TACS) that builds on top of the Attribute Oriented Induction (AOI [4]) algorithm. Oppositely with AOI, our approach performs in an tuple oriented way and aims at reducing the volume of data at the expense of their content and temporal precision. The specificity of this algorithm is also to let the expert decide on how well the temporal ordering of news is preserved during the summarization process. This allows to produce summary precise up to a certain time window.
- We show how frequent mining algorithms (e.g., PrefixSpan) can be applied to the summarized dataset of news and produce useful results. We also show how these results can be refined to obtain a mining at any level of precision with much improved performance when compared with a mining performed directly on the original news descriptions.
- We show how the summary also contributes to the robustness of the algorithm with respect to noisy patterns. In a nutshell, the summary absorbs noisy patterns since these are particularly prone, by their repeating nature, to summarization.

The framework presented here can be extended to account for works in domain such as text processing and information retrieval.

2. MINING FINANCIAL NEWS DATA

Mining frequent patterns in financial news is a challenging task that involves the choices and preferences of domain experts. This is why we designed the *S*treaming *Te*mpor*a*l *D*ata (*STEAD*) analysis framework that lets domain experts express which aspects of the data they are interested in. This user-centric *STEAD* analysis framework is illustrated in Figure 1. It is organized around three services and each service accepts as input, in addition to a dataset, a number of domain specific parameters set by the domain expert:

1. Preprocessing service: This service is responsible for transforming a raw and poorly structured dataset into a *structured* dataset. It consists of (i) "splitting" the news into sequences corresponding to the objects of interest (e.g., if the objects of interest are companies, it will produce, for each company, a distinct sequence of news, all related to that company) and (ii) enriching the news by extracting from the unstructured attributes (e.g., the news story) a set of descriptive features.
2. Summarization service: This service considers as input the structured news and produces a new dataset that is less precise but more concise than the original one. The summarized dataset preserve the structure and the order of tuples of the original dataset. Each summary describes a subset of the original news using an expert defined vocabulary.
3. Mining service: This service identifies the frequent pattern in either the structured news or the summarized news.

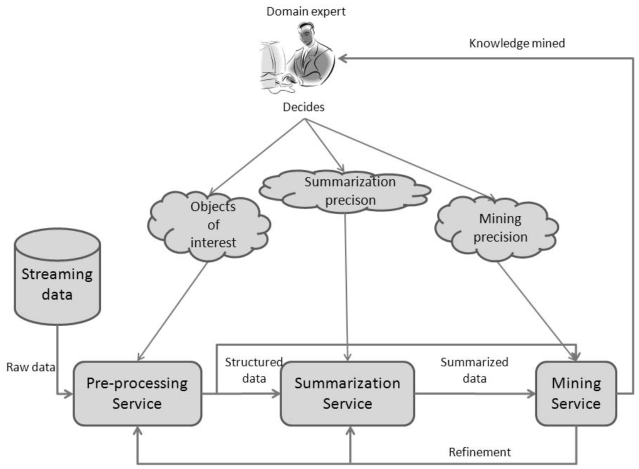


Figure 1: *STEAD* analysis framework for financial news

2.1 News preprocessing service

We mentioned that news semantics is contained in either multi-valued or plain text attributes which are not directly processable for the purpose of frequent pattern extraction. The news preprocessing service is responsible for describing the content of news with categorical attributes. However, due to the very specialized way by which each expert apprehends the news, it is not possible to provide a generic set of descriptive attributes that would satisfy all the experts. Therefore, our preprocessing lets the expert define descriptive attributes and customize, if experts desire so, how attributes are valued from the content of the news.

We base our processing on concept hierarchies. The processing itself can be best explained through an example:

Example 1 (Multi-valued attribute rewriting) *The attribute `topic_code` is a multi-valued attribute used to provide an overall description of the content of a piece of news in the form of a series of codes. Codes can be grouped into themes such as illustrated in the following table:*

Suppose there is a piece of news with a topic code "STX DE GB FR"; it means the news is about stock markets and involves Germany, the UK and France. Using the concept hierarchy presented in Figure 2, this news would be described by two attributes. A first attribute `Location`, valued with the code "WEU", standing for Western Europe and a second attribute `Equitites` with value "STX".

The above example is intended only to show the general process. The algorithm can of course be refined and text

Table 1: Example of codes

Code	Theme	Meaning
STX	Equities	All news about equity markets operations, regulations and structure; additions and deletions from stock indexes
DE	Location	Germany
GB	Location	United Kingdom
FR	Location	France

mining approaches, which have been the object of an important literature, can be leveraged in this context. Algorithm 1 illustrates in more details a naive approach for computing such description. In this algorithm, MSCS(x, y, H) is a function which finds in a given concept hierarchy H (e.g. Figure 2) the most specific common concept that generalizes both x and y. As an example, in Figure 2, MSCS($France$, $Germany$, $H_{Location}$) is *Western Europe* and MSCS($France$, $Russia$, $H_{Location}$) is *Europe*.

Algorithm 1 Data Restructuring (DR) algorithm

```
Input:
    • $inList$ list of multi-valued attribute values for attribute $A_i$
    • $H_{A_i}$ concept hierarchy for attribute $A_i$
Output:
    • $outValue$
Initialize $outValue$ with the first value in $inList$;
for all Attribute value $inValue$ in $inList$ do
    $inList \leftarrow$ MSCS($inValue$, $outValue$, $H_{A_i}$);
    {MSCS : Most Specific Common Subsumer}
end for
return $outValue$
```

Our implementation allows the expert to manually define concept hierarchies or automate the process using WordNet [1] ontologies (as in Figure 2), a screenshot of concept hierarchies generated with WordNet is given in Figure 3.

The processing of plain text follows a similar approach. We use the concept hierarchies designed by the expert (or automatically generated) to identify key words in the news' textual content. Again, more advanced algorithm may be introduced in the *STEAD* analysis framework to identify descriptive keywords of a text and provide a semantic description of it.

2.2 Summarization service

The summarization service provides a semantic compression of the structured dataset produced by the preprocessing service. Summarized news have the same structure as structured news but represent several news, described by more general and less precise descriptors expressed through concept hierarchies.

As we mentioned, financial news are produced in large volumes and mining these volumes to obtain relevant frequent patterns needs to be done at a low level of support. However, direct mining of frequent patterns at low levels

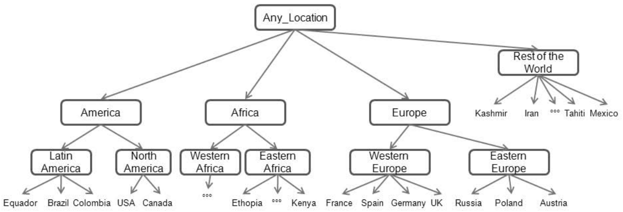
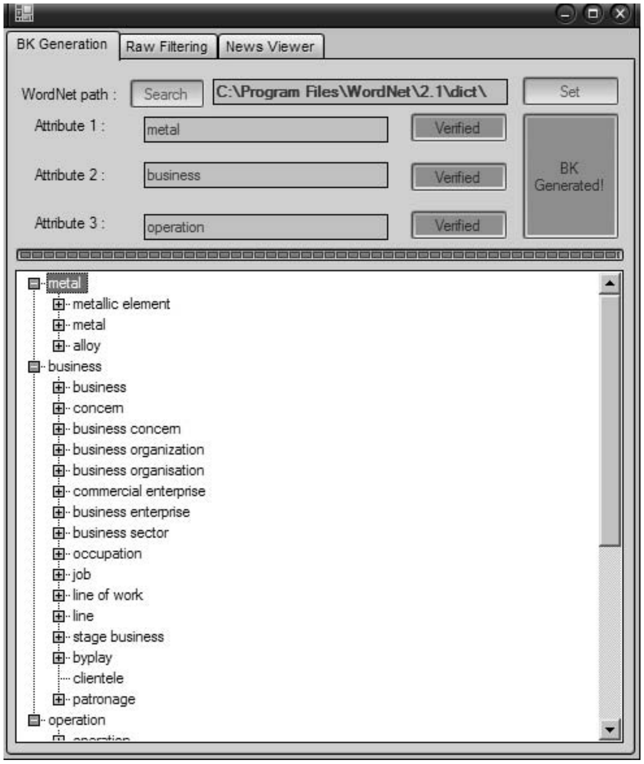


Figure 2: Example of hierarchy of concept for the *Location* attribute



WordNet path : C:\Program Files\WordNet\2.1\dict\

Attribute 1 : metal

Attribute 2 : business

Attribute 3 : operation

metal
- metallic element
- metal
- alloy

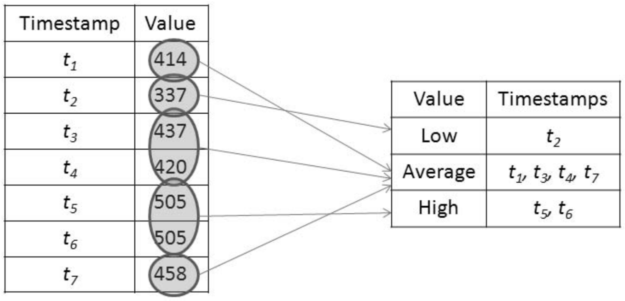
business
- business
- concern
- business concern
- business organization
- business organisation
- commercial enterprise
- business enterprise
- business sector
- occupation
- job
- line of work
- line
- stage business
- byplay
- clientele
- patronage

operation
- operation

Figure 3: Screenshot of concept hierarchies generated using WordNet

of support on these volumes is prohibitively time consuming. By first summarizing this content, experts can discover trends in frequent patterns at high levels of support. These trends can then be refined to arbitrary precision. The total processing time of summarizing, mining and refining is in most cases much lower than the equivalent processing performed on the non-summarized dataset. The exact performance gain depends on the interest expressed. Moreover, shorter processing time is not the only benefit of mining on summarized data. We mentioned that frequent mining performed on the original data would produce a number of noisy patterns. The summarization step reduces the generation of such patterns since redundant news are merged into a single summarized news.

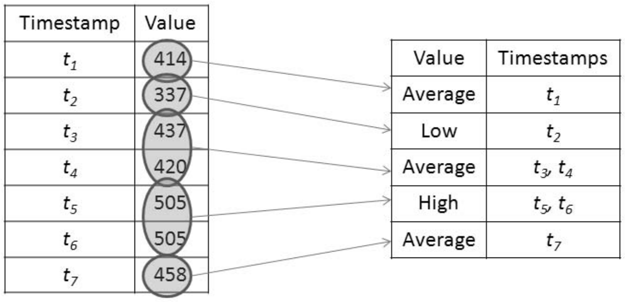
Semantic summarization [8] is performed by grouping together similar news into a single, less precise, summarized news. In order to maximize both the precision of summaries and the final data size, semantic summarization algorithms (Fascicules [6], ItCompress [5], Spartan [3] or SaintEtiQ [8]) typically attempt to group together most similar news as illustrated in Figure 4.



Timestamp	Value
t_1	414
t_2	337
t_3	437
t_4	420
t_5	505
t_6	505
t_7	458

Value	Timestamps
Low	t_2
Average	t_1, t_3, t_4, t_7
High	t_5, t_6

Figure 4: Non-order preserving summary



Timestamp	Value
t_1	414
t_2	337
t_3	437
t_4	420
t_5	505
t_6	505
t_7	458

Value	Timestamps
Average	t_1
Low	t_2
Average	t_3, t_4
High	t_5, t_6
Average	t_7

Figure 5: Order-preserving summary

Frequent pattern mining algorithms rely on the sequentiality of the data to extract patterns. To allow mining of frequent patterns, a summary must preserve—to some extent—the original ordering of news. We therefore introduce a new summarization algorithm that preserve this requirement called the *Order-Preserving (OP)* constraint. We illustrate the OP constraint in Figure 5 and define it as follows : suppose R is a relation defined on attributes $A = \{A_1, A_2, ..., A_n\}$, $n \in N$, and let $\{t_1, t_2, ..., t_m\}$, $m \in N$, be m tuples in relation R. We define $\preceq$ as a total order binary relation over R such as: $\forall(t_i, t_j) \in R^2$ with $i \leq j \leq m$, $arrivalTime(t_i) \leq arrivalTime(t_j) \Leftrightarrow t_i \preceq t_j$.

However, Figure 4 and 5 show that designing a summary that strictly respects the OP constraint can yield poor results as some items are repeated (e.g. Average) and therefore can generate *noisy patterns* during the mining. In order to mitigate this phenomenon, we allow the expert to choose to which extent she wishes to preserve the order. For instance, she may specify that order needs to be preserved to the level of the day or to the level of the week. We call ω the *window of acceptable OP constraint violation* within which the TACS structure respects the OP constraint. ω that can be defined in terms of number of news items (this is the case of the current implementation) or in terms of time period.

Now, TACS is a structure designed in the framework for supporting demanding mining algorithms. The algorithm, inspired by the Attribute Oriented Induction mining algorithm proposed by Han and Fu [4], is a two step Generealize & Merge process. Incoming structured news items t_i are generalized into t_{z_i} on each attribute (e.g. Location, business, etc...) to a specified level (e.g. 1 generalization). For instance, a news about *France* and a news about *Spain* would be described as a news about *Western Europe* : Generalization step.

Then, each generalized news t_{z_i} will be compared to previously generalized news t_{z-last_j} within the window ω. If t_{z_i} is identical to a t_{z-last_k} then a COUNT attribute on the number of news items represented by t_{z-last_k} will be updated and a pointer toward t_{z_i} will be stored. In the case t_{z_i} is not identical to any t_{z-last_k}, t_{z_i} will be added to the summary table with COUNT = 1 and the windows ω will be slid forward : Merging step.

The output of this process is an OP summary that complies to the OP constraint within a sliding window of size ω. We give a screenshot of the output of our prototype with $\omega = 5$ performed over a month worth of real world financial news in Figure 6.

Therefore, the expert can express the overall level of precision of the TACS structure by indicating for each attribute its level of generalization. By default, our implementation fixes this level to 1. However, the expert can chose to give more importance to an attribute by lowering the level of generalization (e.g. 0) or increase the level (e.g. 3 or 4) to diminish the importance of the attribute.

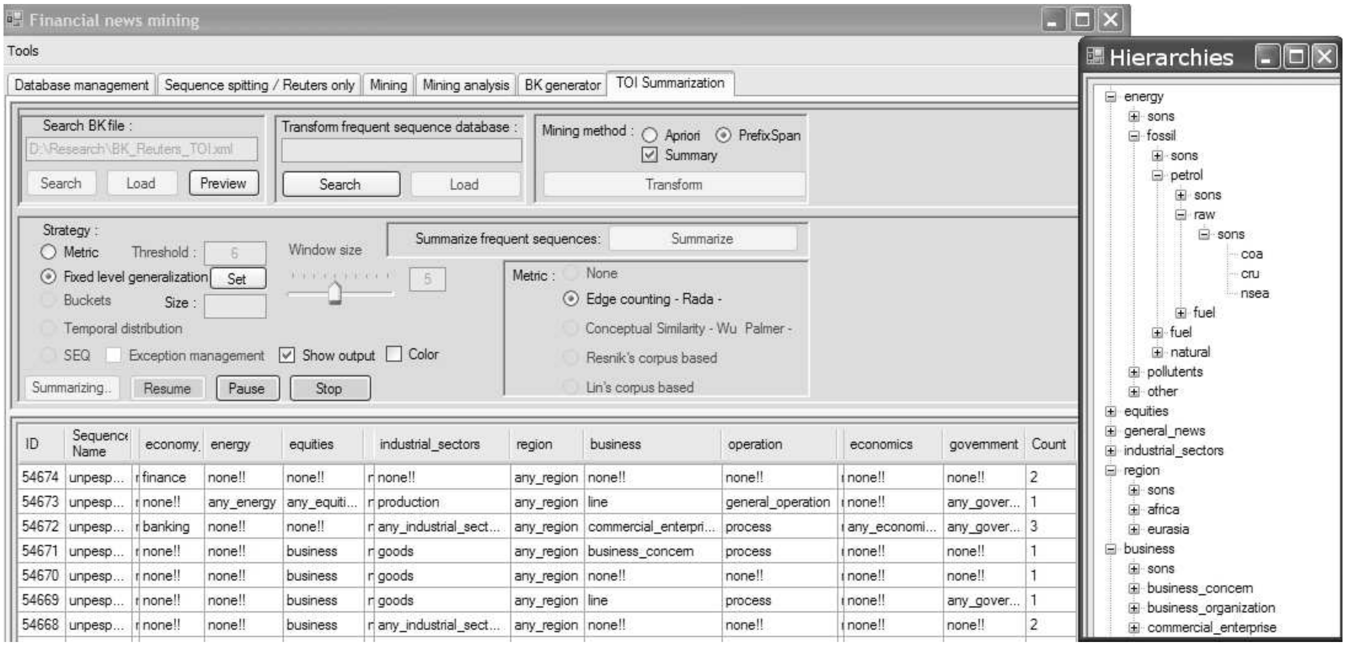
2.3 Frequent pattern mining service

The frequent pattern mining service processes the summarized news to discover frequent patterns. An algorithm for frequent pattern mining was first proposed by Agrawal and Srikant in 1995 [2] with an Apriori-based approach. Since then, more advanced algorithms (e.g. GSP, SPADE, PrefixSpan [7], SPAM, etc...) have been proposed to mitigate Apriori-based algorithms' weaknesses which are the possible combinatory explosion at candidate generation time and the multiple scans over the data. We based our implementation on Pei et al.'s PrefixSpan approach. This approach reduces the number of candidates by using projected databases. The only parameter required from the expert for the mining algorithm to perform is the desired *support* of frequent patterns.

The patterns mined over the summarized news are less precise than when the mining is performed on the structured news. However, the summarization allows to achieve mining of patterns with a higher level of support while reducing the number of noisy patterns.

3. DEMONSTRATION

Our demonstration consists of illustrating the three main



Search BK file : D:\Research\BK_Reuters_TOI.xml

Transform frequent sequence database :

Mining method : ○ Apriori ● PrefixSpan
☑ Summary

Strategy :
○ Metric Threshold : 6
● Fixed level generalization
○ Buckets Size :
○ Temporal distribution
○ SEQ □ Exception management ☑ Show output □ Color

Window size 5

Summarize frequent sequences:

Metric : ○ None
● Edge counting - Rada -
○ Conceptual Similarity - Wu Palmer -
○ Resnik's corpus based
○ Lin's corpus based

ID	Sequence Name		economy	energy	equities		industrial_sectors	region	business	operation		economics	government	Count
54674	unpesp...	r	finance	none!!	none!!	r	none!!	any_region	none!!	none!!	(	none!!	none!!	2
54673	unpesp...	r	none!!	any_energy	any_equiti...	r	production	any_region	line	general_operation	(	none!!	any_gover...	1
54672	unpesp...	r	banking	none!!	none!!	r	any_industrial_sect...	any_region	commercial_enterpri...	process	(	any_economi...	any_gover...	3
54671	unpesp...	r	none!!	none!!	business	r	goods	any_region	business_concern	process	(	none!!	none!!	1
54670	unpesp...	r	none!!	none!!	business	r	goods	any_region	none!!	none!!	(	none!!	none!!	1
54669	unpesp...	r	none!!	none!!	business	r	goods	any_region	line	process	(	none!!	any_gover...	1
54668	unpesp...	r	none!!	none!!	business	r	any_industrial_sect...	any_region	none!!	none!!	(	none!!	none!!	2

```
⊟ energy
    ⊞ sons
    ⊟ fossil
        ⊞ sons
        ⊟ petrol
            ⊞ sons
            ⊟ raw
                ⊟ sons
                    coa
                    cru
                    nsea
            ⊞ fuel
        ⊞ fuel
        ⊞ natural
    ⊞ pollutents
    ⊞ other
⊞ equities
⊞ general_news
⊞ industrial_sectors
⊟ region
    ⊞ sons
    ⊞ africa
    ⊞ eurasia
⊟ business
    ⊞ sons
    ⊞ business_concern
    ⊞ business_organization
    ⊞ commercial_enterprise
```

Figure 6: Screenshot of TACS over real world financial news and concept hierarchies used

steps of our framework for mining financial news issued by real world news sources.

Financial news pre-processing. Financial news issued by media companies such as Reuters are very unstructured and poorly exploitable data. We show how we split the global worldwide stream of news into news time sequences for companies. We then present show how experts can preprocess such data by generating concept hierarchies using WordNet for restructuring and enriching the news representation. The set of features of interest defined earlier will allow to transform multi-valued attributes into mono-valued attributes and extract additional features from the text corpus.

TACS summarization service. Our TACS process allows much flexibility with different choices for generalizing and merging the news. We will show how financial news data can be summarized incrementally, in an order-preserving fashion and also the benefit of some parameters such as loosening the OP constraint.

Mining frequent patterns on summaries. The topology of TACS reflects the structure of raw news even after all the preprocessing work. It was designed to be exploitable as it is, without the need for a *decompression* step. Therefore, we will show how a conventional mining algorithm such as PrefixSpan performs seamlessly on both preprocessed and summarized news items. The results on both data sets will of course be different, but these results will allow us to open a discussion the benefits of our TACS approach for supporting such applications.

4. ACKNOWLEDGEMENTS

This work was supported by the ADAGE project from the Information Systems, Technology and Management School and the Service Oriented Architecture (SOA) group from the School of Computer Science and Engineering, UNSW, Sydney, Australia. The authors would also like to thank SIRCA for providing them with the news data and the reviewers for their suggestions and advices.

5. REFERENCES

[1] Wordnet. http://wordnet.princeton.edu/.

[2] R. Agrawal and R. Srikant. Mining sequential patterns. In *Proc. of the 11th International Conference on Data Engineering (ICDE 1995)*, 1995.

[3] S. Babu, M. Garofalakis, and R. Rastogi. Spartan: A model-based semantic compression system for massive data tables. In *Proc. of the ACM SIGMOD International Conference on Management of Data (SIGMOD 2001)*, 2001.

[4] J. Han and Y. Fu. Exploration of the power of attribute-oriented induction in data mining. *Advances in Knowledge Discovery and Data Mining*, 1996.

[5] H. Jagadish, R. Ng, B. Ooi, and A. Tung. Itcompress: an iterative semantic compression algorithm. In *Proc. of the 20th International Conference on Data Engineering (ICDE 2004)*, Mar 2004.

[6] H. V. Jagadish, J. Madar, and R. T. Ng. Semantic compression and pattern extraction with fascicles. In *Proc. of the 25th International Conference on Very Large Databases (VLDB 1999)*, 1999.

[7] J. Pei, J. Han, B. Mortazavi-Asl, and H. Pinto. Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth. In *Proc. of the 17th International Conference on Data Engineering (ICDE 2001)*, 2001.

[8] R. Saint-Paul, G. Raschia, and N. Mouaddib. General purpose database summarization. In *Proc. of the 31st International Conference on Very Large Databases (VLDB 2005)*, pages 733–744, August 2005.