Technique for Reducing the Number of Rules in a Temporal Knowledge Base

Maria Antonina Mach[1], Pawel J. Kalczynski[2]

[1]University of Economics, Dept. of Artificial Intelligence Systems
Komandorska 118/120, 50-345 Wroclaw, Poland
maria.mach@ae.wroc.pl

[2]University of Toledo, College of Business Administration, MS#103
Dept. of Information, Operations & Technology Management 2801 West Bancroft St.
Toledo, OH 43606
pawel.kalczynski@utoledo.edu

Abstract

A knowledge base about any domain should satisfy some basic properties. One of these properties is completeness. Assuming that a knowledge base is temporal, that is, it contains explicit time references in its rules and facts, we may face a problem of an infinite number of possible premises, as the number of temporal references (e.g. in business documents) is unlimited. A technique for reducing the number of possible rules in a temporal knowledge base is proposed in this paper. This is achieved by transforming an infinite set of possible temporal rules into a finite one by mapping an infinite set of temporal references onto a finite set, and by automatic trimming of idiomatic phrases read from business documents. The technique for reducing the number of rules is presented and illustrated by an example.

1. Introduction

The problem of knowledge base reduction emerged from a major research project on temporal semantic networks. The main objective of the former project was to develop a time-aware structure to represent business information for the purpose of analyzing barriers to entry and their changes over time. Barriers to entry were chosen as a sample element of the business environment. A barrier to entry is everything that makes an entry to a branch or a market space difficult for economic entities (OFT, 1994). Gilbert, in turn, defines a barrier to entry as "a rent that is derived from incumbency" (Gilbert, 1989). Just as other components in the business environment, barriers change in time, differ in type (e.g. they are qualitative or quantitative), and affect organizational strategies of doing business.

Therefore, any findings based on this example may be easily extended to other elements of the business environment.

The structure to represent business information should be automatically built based on business information continuously published by trusted sources on the Web. While working on sample temporal rules for capital barriers to entry, we observed that the number of temporal references in business documents is extremely large and, perhaps, even infinite. Therefore, it was necessary to convert the infinite set into a finite one. This conversion enables building a finite set of temporal rules, and thus, obtaining a complete knowledge base. The main problem was how to perform the conversion operation. We also observed that many idiomatic phrases differ only slightly (e.g. "sudden increase of oil prices" and "increase of oil prices"), therefore the number of possible non-temporal natural-language mappings used in rules may be diminished. This observation led to elaborating the second part of the rules-reduction technique.

The remainder of the paper is organized as follows: in Section 2 the motivation for reducing the number of rules in a temporal knowledge base is presented. Section 3 contains a sample set of rules describing capital barriers to entry. In Section 4 a technique for reducing the number of possible temporal rules is described followed by the concluding remarks and future research ideas.

2. Motivation

A knowledge base is one of the elements of the temporal reasoning system that uses business news archives to analyze capital barriers to entry. The system uses a time-aware structure (temporal semantic web) that stores business knowledge extracted from business news available on the Web. The rule base for this system is stored in a separate structure; thus the knowledge base consists of two physical constructs. The temporal semantic web is built automatically by parsing and indexing a large corpus of business documents. Sowa's conceptual graphs (Sowa, 2000) extended with temporal annotations (Moulin, 1997) serve as the knowledge representation model. The rule base, on the other hand, is prepared by a knowledge engineer co-operating with experts. The rules should reflect relationships among events and features that can be found in business documents, in order to provide an appropriate analysis. Therefore, rules' components are usually parsed from business documents, in the form of natural language mappings (e.g. "a sudden increase of oil prices has been observed lately").

Let us consider one example of capital barriers to entry. Two economic variables that influence the height of capital barriers to entry have been chosen, namely, (1) the U.S. dollar exchange rate and (2) oil prices in U.S. dollars (denoted DER and OP, respectively). One must notice, however, that these are not the only variables responsible for capital barriers; they were chosen as there is much information about them available in business documents. Because time is considered the basic dimension in temporal reasoning, each rule has three

components in its premises: (1) a temporal reference (denoted as *Tref*), (2) DER, and (3) OP. The *Tref* variable is necessary for the inference engine to make use of temporal references extracted automatically from business documents. The rules are described in more detail in the next section.

It should be noted that, in reality, the number of temporal rules built for the corpus of business documents is unlimited, as the number of temporal expressions in the documents is practically unlimited. This, in turn, leads to an observation that has been almost impossible to build a knowledge base that will satisfy all the basic properties, such as consistency, completeness, no subsumed or redundant rules etc. (Ligęza, 2005). The completeness property has been particularly hard to satisfy.

In response to this problem this paper presents a technique that would allow reducing the number of possible rules from infinity into a finite set.

3. Illustration of the problem: Dollar Exchange Rate and Oil Price in Time

In the process of formulating sample rules for reasoning about capital barriers to entry, it has been assumed that the premises of each rule contain three elements (variables): *Tref*, DER and OP, which stand for a temporal reference, dollar exchange rate, and oil price, respectively. According to this assumption, the rules may have the following forms:

1. IF at Tref (description) DER (description) THEN conclusion
2. IF at Tref (description) OP (description) THEN conclusion
3. IF at Tref (description) DER (description) AND OP (description) THEN conclusion
4. IF at Tref (description) DER (description) OR OP (description) THEN conclusion

Next, a list of sample values of each variable in rules' premises was prepared. The values of *Tref* were computed based on the TREC/FT collection of over 200,000 Financial Times documents, while the values of OP and DER (in the form of natural language mappings) came from an ontology. The latter was created based on a literature review (Bain, 1993). These sample values are presented in **Table 1**.

Given the instances of rules and the values of variables in rules in Table 1, the knowledge base would contain 50,622 rules. In addition, some of the rules may be contradictory, but this problem is outside the scope of this paper. The number of rules was calculated using the decision table theory. A decision table is a tabular form that presents a set of conditions and their corresponding actions (Metzner & Barnes, 1977). The formula for computing the maximum number of rules is given by:

$$mnr = m \times n \,. \qquad (1)$$

where:

mnr – the maximum number of rules
m – the number of conditions in rules' premises
n – number of possible values of conditions.

Using this method, and having 39 temporal references, 26 mappings (values) of DER and 24 mappings (values) of OP (see Table 1), we obtain:

1. Tref + DER: 39 * 26 = 1,014
2. Tref + OP: 39 * 24 = 936
3. Tref + DER + OP (conjunction of DER and OP): 39 * 26 * 24 = 24,336
4. Tref + DER + OP (disjunction of DER and OP): 39 * 26 * 24 = 24,336

which together comes to 50,622 possibilities (rules).

Because the number of temporal references in business news is infinite, this approach would render the knowledge base incomplete.

Table 1. Values of Variables in the Sample Capital Barrier to Entry.

TEMPORAL REFERENCE (TREF)	TEMPORAL REFERENCE (TREF) – cont.	DOLLAR'S EXCHANGE RATE (DER)	OIL PRICES (OP)
Past:	Pointer:	Sudden increase of DER	Sudden increase of OP
Over the previous three months	As of May 2002	Rapid increase of DER	Quick increase of OP
Some time ago	In the first quarter of 2004	Significant increase of DER	Significant increase of OP
Last week	On May 3, 2005	Insignificant increase of DER	Insignificant increase of OP
On the recent Monday	On Tuesday	Unexpected increase of DER	Unexpected increase of OP
Yesterday	In June	Previously unexpected increase of DER	Previously unexpected increase of OP
A few days ago	Parts:		
Presence:	At the beginning of/start of July	Sudden decrease of DER	Sudden decrease of OP
In the current month	Since the end of 2004	Rapid decrease of DER	Quick decrease of OP
In the present quarter	In mid/the middle of July	Significant decrease of DER	Significant decrease of OP
This year	Early/late in July	Insignificant decrease of DER	Insignificant decrease of OP
Today	Intervals/operators:	Unexpected decrease of DER	Unexpected decrease of OP
Now	From June 16 to July 10	Previously unexpected decrease of DER	Previously unexpected decrease of OP
Future:	Between March and May		
In the month to come	In five weeks	Low DER	Low OP
In the forthcoming elections	Two weeks after Christmas	High DER	High OP
Will take place on Monday	Three days later	Stable DER	Stable OP
Next week	Five months before/back/earlier	Unstable DER	Unstable OP
Soon			
Over the next three days		Sudden changes of	Sudden changes of OP
Tomorrow			Quick changes of OP
Continuity:			
Since/from the end of June			
Over/during/for/within the next few days			
In the past three			

weeks By/until/till the end of March 2003 The following three weeks		DER Quick changes of DER Significant changes of DER Insignificant changes of DER Unexpected changes of DER Previously unexpected changes of DER Frequent changes of DER Frequent fluctuations of DER Strengthening of the dollar Weakening of the dollar	Significant changes of OP Insignificant changes of OP Unexpected changes of OP Previously unexpected changes of OP Frequent changes of OP Frequent fluctuations of OP

4. Technique for Reducing the Number of Rules

The first part of our technique consists of transforming any temporal expression extracted from a business document (e.g. "on Dec. 31, 2005," "last week," "today"), into a reference from a finite set. In this way, the number of possible temporal elements in rules' premises would be finite.

People perceive time and temporal elements as fuzzy notions (Kalczynski & Chou, 2005). Therefore, it is assumed, that there exists a finite set of temporal fuzzy references, into which any temporal reference, either strong or weak, may be transformed by means of a specific function. The notions of strong and weak temporal references are defined in (Abramowicz, Kalczynski, & Wecel, 2002). So, the first part of our technique consists of a finite transformation of in-text temporal references.

4.1. Reducing the Number of Possible Temporal References

We assume that time is granular (not continuous) in nature. Formally, a granularity *G* partitions a granularity *H* if it is finer than *H* (Bettini, Jajodia, & Wang, 2000) and it groups into *H* (Bettini et al., 2000). In other words, a granularity *G* partitions *H* if any granule of *H* can be represented as a union of granules of *G*. For example, any given year, quarter, month or week can be represented as a union of days. Therefore *DayGty* partitions *YearGty*, *QuarterGty*, *MonthGty* and *WeekGty* respectively. For more detailed formal descriptions of temporal granularities one can see (Bettini et al., 2000) and (Bettini & Ruffini, 2003) for example.

We assume the existence of the finest temporal granularity *B* (e.g. *DayGty*) that partitions any other granularity in the system. Therefore, *B* can serve as a referential set for fuzzy representations of granules in the granularity system. Any temporal granule in the system can be represented as a mapping $\underline{F}$: B→[0,1] that shows the degree of membership of granules of *B* in the granule in question (because *B* partitions any other granularity in the system). This mapping is compatible with a common definition of fuzzy sets; see (Dubois & Prade, 2000) for example. In addition, the fuzzy representation extends (not substitutes) the traditional notion of temporal granularities. Therefore, the representation also works for precise expressions such as "today," or "on Dec. 24, 2005."

We define a function *TIndex* that takes a temporal expression *e*, temporal context *c*, and referential set *B* as arguments and returns a fuzzy set of granules $\underline{F}$ representing the expression *e* in terms of *B*. Formally:

$$TIndex(e,c,B) = \underline{F} : B \rightarrow [0,1] . \quad \textbf{(2)}$$

In theory, *TIndex* can represent any temporal expression in the natural language in terms of the referential set, provided that the temporal context, i.e., frame of reference, is given. For example, "last month," "today," "three years ago" can be represented in terms of *B* if the publication date of the source document is known. One may wonder how the values of the membership function are assigned to $\underline{F}$.

An extension of the Time Indexer (Kalczynski, Abramowicz, Wecel, & Kaczmarek, 2003) allows representing most temporal expressions used in business news documents as sets of granules. The Natural Language Processing (NLP) techniques are used to extract the temporal context. The values of the membership function are obtained from user studies such as (Kalczynski & Chou, 2005). The representations of common uncertain expressions show certain patterns and can be generalized and approximated (Kalczynski & Chou, 2005). In this way, it is now possible to represent such expressions like, for example, "on Dec. 24, 2005", "yesterday," "in May," "last year," "two weeks ago," "since the beginning of July," "by June," "between June and August," "recently," or "currently."

Even if we assume that the referential set *B* is finite, the number of natural expressions referring to time is unimaginably large. However, the concept of inclusion from the Fuzzy Set Theory seems to be useful for building a reduced temporal representation.

Assume the following set of seven arbitrarily selected natural language expressions describing time relative to some frame of reference:

T={in the distant past, in the past, recently, currently, soon, in the future, in the distant future}

One can observe that, although it seems that the set could be ordered in terms of time, it can neither be fully nor partially ordered. Once the elements of *T* are represented in terms of the base granularity *B*, they will not contain distinct

granules. What is more, each of these temporal expressions is fuzzy. When converted to the base granularity with the *TIndex* function, the elements of *T* will be transformed into the following fuzzy sets: $\underline{E}_{-3}$, $\underline{E}_{-2}$, $\underline{E}_{-1}$, $\underline{E}_{0}$, $\underline{E}_{1}$, $\underline{E}_{2}$, and $\underline{E}_{3}$. **Figure 1** illustrates the concept of fuzzy representations of elements of the set *T*.

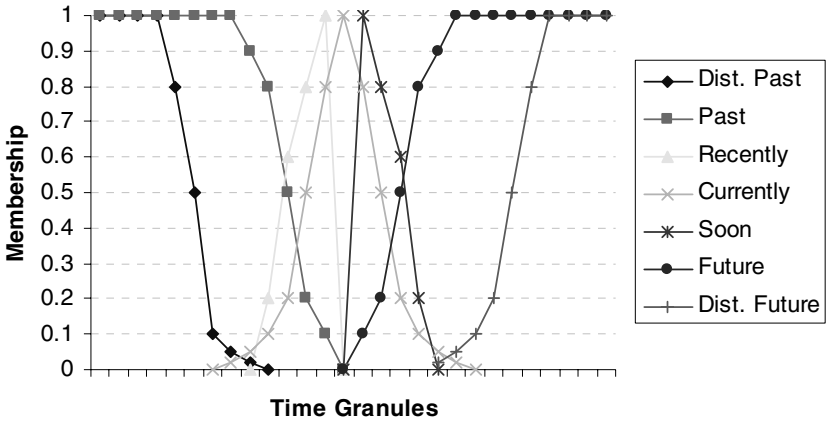

Figure 1. Fuzzy Representations of Temporal Expressions in T.

Similarly, each temporal expression extracted from the document content is transformed with the *TIndex* function to its granular representation. As a result one gets a fuzzy set $\underline{F}$ of granules representing the temporal expression in terms of the finest assumed granularity *B*.

Next, one of the fuzzy set comparison indexes can be used to determine the equality between the representation of the reference in question ($\underline{F}$) and the fuzzy sets representing the elements of *T* ($\underline{E}_{-3}$, $\underline{E}_{-2}$, $\underline{E}_{-1}$, $\underline{E}_{0}$, $\underline{E}_{1}$, $\underline{E}_{2}$, and $\underline{E}_{3}$). The equality between fuzzy sets is typically measured by the degree to which both $\underline{F}$ is included in $\underline{E}_i$ and $\underline{E}_i$ is included in $\underline{F}$ (Dubois & Prade, 2000). Therefore the following relative equality index will be used to compare fuzzy sets of granules (Dubois & Prade, 2000):

$$REC\left(\underline{F},\underline{E}\right)=\frac{\left|\underline{F}\cap\underline{E}\right|_F}{\left|\underline{F}\cup\underline{E}\right|_F}, \qquad (3)$$

where $|\underline{D}|_F$ denotes the cardinality of a fuzzy set $\underline{D}$ computed as $\sum_{b\in B}\underline{D}(b)$ (De Luca & Termini, 1972) and the definition of union and intersection of fuzzy sets can be found in the classical work of Zadeh (Zadeh, 1965).

Let *e* be a temporal expression in the document content such as "lately," "in January," or "on Dec 1, 2005". Let *c* be a granule of certain granularity representing the frame of reference (context) for temporal expression *e* (e.g. the publication date of the analyzed document). Let *B* be the finest granularity in the system (e.g. *Days*). Further let t_i be any element of *T*. We define a function *TTF*

that transforms a temporal reference *e* into the index value of one of the elements in *T*. The function is given by:

$$TTF(e,c,B,T) = i : \max_i \left(REC(\underline{F}, E_i) \right), \tag{4}$$

where $\underline{F} = TIndex(e,c,B)$, and $\underline{E}_i = TIndex(t_i,c,B)$.

Observe that the choice of elements for *T* is arbitrary. However, it is reasonable to select significant elements which enable comparing reduced temporal references with the desired precision.

All fuzzy sets of granules can be represented in terms of the relational model, thus they can be stored in a relational database (Kalczynski, forthcoming). All computations necessary to achieve the reduced representation may be done using the Structured Query Language (SQL).

As an example consider a fictitious document published, say, on Nov. 1, 2005 with the following three expressions in the content:

1. "…the sudden increase of oil price that happened three months ago."
2. "… we will know within the next two days"
3. " is expected to decrease on Nov 7."

If we use the previously assumed set *T* for reduction, the first reference will become "distant past," the second will turn to "soon," and the last one will be classified as "future." By analogy, any other reference in the document will be classified as one of the elements of *T*.

4.2. Reducing the number of natural language mappings

The second part of our technique enables further reduction of the number of rules in the knowledge base by limiting the number of possible values of linguistic variables DER and OP; that is, by reducing the number of in-text natural language mappings. Consider an example taken from Table 1:

Sudden increase of DER
Rapid increase of DER
Significant increase of DER
Insignificant increase of DER
Unexpected increase of DER
Previously unexpected increase of DER

All the above mappings describe an increase of DER. Each of these mappings is accompanied by information about a particular characteristic of this increase. Nevertheless, what is most important in each of these mappings is the *increase* of DER. Therefore, by trimming these statements we can preserve the essential information and reduce the number of possible variable values at the same time.

Generally, three types of linguistic structures may be distinguished in the above list of mappings:

1. *[adverb] [adjective] noun$_1$ preposition noun$_2$*

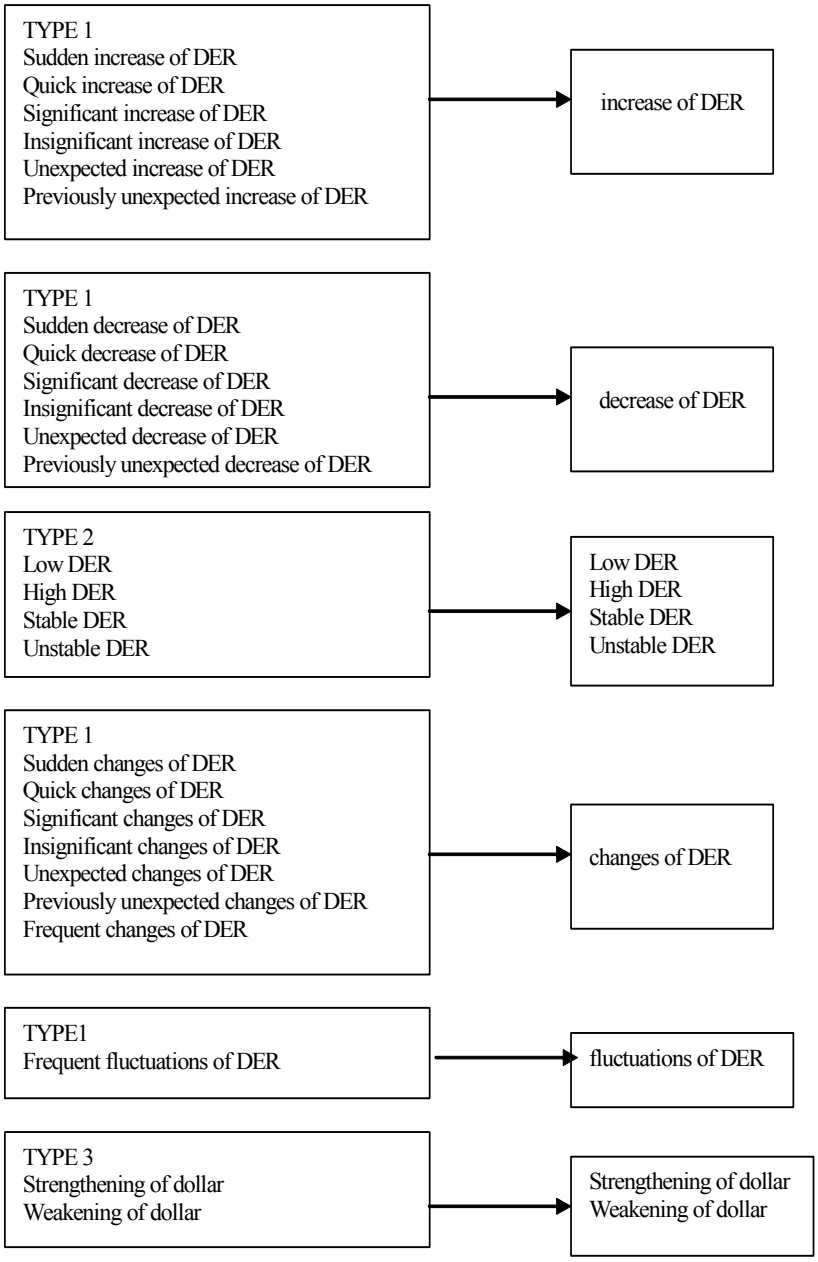

Figure 2. Transformation of DER descriptions.

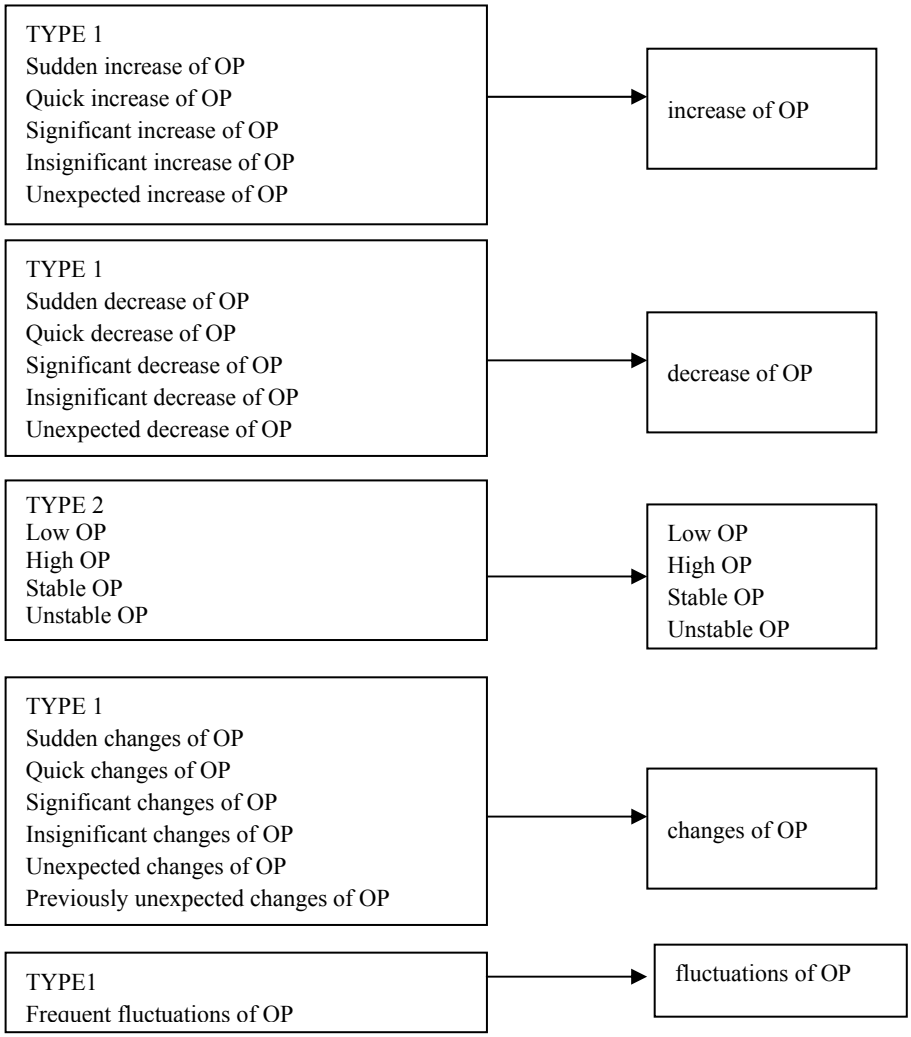

Figure 3. Transformation of OP descriptions.

2. *adjective noun*
3. *gerund preposition noun*

In order to reduce the number of possible mappings, the optional part may be trimmed by the system, according to the following algorithm:

1. Group idiomatic expressions according to their structure (Type 1, 2 or 3)
2. For every Type 1 structure, eliminate adverbs and adjectives using a built-in dictionary, thus simplifying the structure of the form: $noun_1$ *preposition* $noun_2$

3. Leave Type 2 and Type 3 structures unchanged.

By transforming the linguistic values of OP and DER presented in **Table 1** according to the above algorithm one obtains a reduced list of ten values of the linguistic variable DER and eight values of the linguistic variable OP. The initial and reduced lists of variable values are presented in Figures 2 and 3.

Using the previously assumed set *T* of seven temporal references, ten values (mappings) of DER and eight values (mappings) of OP (see Figures 2 and 3), the initial set of 50,622 reduces to:

1. Tref + DER: 7 * 10 = 70
2. Tref + OP: 7 * 8 = 56
3. Tref + DER + OP (conjunction of DER and OP): 7 * 10 * 8 = 560
4. Tref + DER + OP (disjunction of DER and OP): 7 * 10 * 8 = 560

which comes to a total of 1,246 possibilities (rules). Compared to the initial sample rule base the number of rules was reduced by 97.54%.

5. Conclusions

The present paper offers a description of a new technique for reducing the number of temporal inference rules in a temporal knowledge base. This reduction is necessary for building temporal inference rules based on business documents. In reality, the number of possible temporal references in business documents (and thus in rules) is infinite and this leads to the infinity problem in the knowledge base. Without reducing the number of possible rules, a (temporal) knowledge base would never be complete.

The main advantage of the presented technique is that it enables building a finite, operating knowledge base on capital barriers to entry, or potentially on any other business domain. The knowledge base may represent information found in business documents and, at the same time, preserve its temporal aspect. The trade-off of building such knowledge base is a slight loss of precision of information. Because the Type 1 idiomatic expressions are trimmed, the representational power of rules is lower. Nevertheless, the essential information (e.g. "increase of DER") is preserved.

As we stated earlier, capital barriers to entry are influenced by many more factors than those used in our illustration. This makes the method of reducing the possible number of variables' values even more useful. Recall that, having only three variables in rules' premises, our technique allows reducing the knowledge base by more than 97%. There are also other barriers to entry; see e.g. (Bain, 1993) for details.

Constructing a working knowledge base for this domain could potentially enable building a temporal intelligent system. Such a system would be capable of reasoning about conditions of entry, and the latter are an important factor

influencing organizational strategy. In our opinion, a temporal intelligent system is needed to capture the complex aspects of economic environment and to analyse them. The pace of change in the economic environment evokes a need to take into consideration the temporal aspect of the environment in an explicit way. A temporal intelligent system would be helpful in such tasks, as: providing an appropriate description of different aspects of the environment, taking into account their temporal characteristics, and unifying those descriptions. This, in turn, would further allow for more general inferencing, historical analyses of changes in the environment and forecasting them.

The main advantage of building and using a temporal intelligent system is related ot the concept of ”the economy of speed” (Tvede & Ohnemus, 2001). The sooner changes in the environment are identified, the sooner strategic adjustment decisions can be made.

The elements of the finite set *T* of temporal references were arbitrarily chosen. And so, future work will focus on verifying the TTF function against human perception of time in business news. The TREC/FT corpus will act as a source of temporal expressions.

6. References

1. Abramowicz, W., Kalczynski, P. J., & Wecel, K. A. (2002). *Filtering the Web to Feed Data Warehouses*. London: Springer-Verlag.
2. Bain, J. S. (1993). *Barriers to new competition : their character and consequences in manufacturing industries*. Fairfield, NJ: A.M. Kelley.
3. Bettini, C., Jajodia, S., & Wang, S. X. (2000). *Time Granularities in Databases, Data Mining, and Temporal Reasoning*. Berlin Heidelberg: Springer-Verlag.
4. Bettini, C., & Ruffini, S. (2003). Direct Granularity Conversions among Temporal Constraints. *Journal of Universal Computer Science, 9*(9), 1123-1136.
5. De Luca, A., & Termini, S. (1972). A Definition of a Nonprobabilistic Entropy in the Setting of Fuzzy Sets Theory. *Information & Control, 20*(4), 301-312.
6. Dubois, D., & Prade, H. M. (2000). *Fundamentals of fuzzy sets*. Boston: Kluwer Academic.
7. Gilbert, R. J. (1989). Mobility Barriers and the Value of Incumbency. In R. Schmalensee & R. D. Willig (Eds.), *Handbook of Industrial Organization* (Vol. 1, pp. 475-535). Amsterdam; New York, N.Y., U.S.A.: North-Holland; Sole distributors for the U.S.A. and Canada, Elsevier Science Pub. Co.
8. Kalczynski, P. J. (forthcoming). Time Dimension for Documents in the Knowledge Warehouse. *Journal of International Technology and Information Management*.
9. Kalczynski, P. J., Abramowicz, W., Wecel, K. A., & Kaczmarek, T. (2003, May 18-21, 2003). *Time Indexer: A Tool for Extracting Temporal References from Business News.* Proceedings of the 2003 Information Resource Management Association International Conference, Philadelphia, PA, pp. 832-835.
10. Kalczynski, P. J., & Chou, A. (2005). Temporal Document Retrieval Model for Business News Archives. *Information Processing & Management, 41*(3), 635-650.
11. Ligęza, A. (2005). *Logical Foundations for Rule-Based Systems*. Krakow, Poland: Uczelniane Wydawnictwa Naukowo-Dydaktyczne Akademii Górniczo-Hutniczej w Krakowie.

12. Metzner, J. R., & Barnes, B. H. (1977). *Decision table languages and systems*. New York: Academic Press.
13. Moulin, B. (1997). Temporal Contexts for Discourse Representation: An Extension of the Conceptual Graph Approach. *Applied Intelligence, 7*(3), 227-255.
14. OFT. (1994). *Barriers to Entry and Exit in UK Competition Policy*: A Report by Office of Fair Trading (OFTRP2).
15. Sowa, J. F. (2000). *Knowledge representation : logical, philosophical, and computational foundations*. Pacific Grove: Brooks/Cole.
16. Tvede, L., & Ohnemus, P. (2001). *Marketing strategies for the new economy*: John Wiley & Sons, Ltd.
17. Zadeh, L. A. (1965). Fuzzy Sets. *Inform. Control, 8*, 338-353.