TEMPORAL INFORMATION INDEXING MODEL

Witold Abramowicz and Andrzej Bassara
Department of Information Systems, Poznan University of Economics, ul. Niepodleglosci 10, Poznan, Poland

Keywords: Information Retrieval, Temoral Information Retrieval, Temporal Expressions, Indexing, Temporal Indexing.

Abstract: Modern information retrieval models are not capable of resolving queries containing temporal criteria. One is not able to search for documents which content relates to certain time (for instance ,,find all documents related to the third quarter of the last year“). This limitation is mainly due to syntactic nature of modern information retrieval models, which perform query-document matching based on syntactic or simplified semantic similarity measures. In this article, we are focusing on the problem of creating document indexes, which represent time to which document contents relate, and which in turn allow for searching documents using temporal criteria.

1 INTRODUCTION

Modern information retrieval (IR) systems are not capable of searching for documents which contain information related to a specified time. It is relatively easy to find documents based on their publication date. Nevertheless, the publication date may be significantly different from the time to which the article relates. Similarly to the bi-temporal databases (Jensen and Snodgrass, 2006), two orthogonal dimensions of time exists: the transaction time and the valid time. The transaction time is specific for a publication process and may include: creation, approval, publication or modification dates. The valid time is the time to which information presented in the article relates.

This limitation is mainly caused by simplification of indexing. Documents are usually indexed automatically with uncontrolled vocabulary. In such case, indexing terms are usually words extracted from document content. Computation of relevance is than based purely on syntactic features. Sometimes words are stemmed or lemmatized. The comparison of query/document terms may by also supported by thesauruses or performed on ontological level. both approaches brings the process closer to semantic level.

Table 1: Sample query with temporal criteria.

Document:	The board of the Globe Trade informs that during 16 August 2006
Information need:	all documents that relate to the third quarter of the last year
Query:	the third quarter of the last year

This approach is, however, not appropriate for queries with temporal criteria. The table 1 presents a sample scenario. It appears that the query and the document are not syntactically similar. The semantic comparison based on concepts comparison will also yield no similarity. The document seems however to be partially relevant. Limiting our consideration only to calendar expressions, the computation of relevance requires:

1. extraction of temporal features from the document and the query – ,,16 August 2006“, ,,the third quarter of the last year (2006)”,
2. encoding their value using a formal time model – `Y2006M08D16`, `Y2006Q3`,
3. comparing the values by means of arithmetic specific for selected time model – `Y2006M08D16` is within `Y2006Q3`,
4. computing the relevance – the references are expressed on different granularity levels (days and quarters), and although one reference contains another, it is not clear how to compute relevance, as it may be dependent on information need.

Successful application of this approach requires, however, addressing following issues:

Time Models Multiplicity. Temporal expressions may be formalized in various time models (point-based, interval-based, point-interval based). These models are also often extended to support multiple time units and imprecise expressions. Moreover, it is not be possible to compare two temporal expressions, unless they are expressed in comparable and known time models.

Abramowicz W. and Bassara A. (2008).
TEMPORAL INFORMATION INDEXING MODEL.
In *Proceedings of the Tenth International Conference on Enterprise Information Systems - AIDSS*, pages 387-390
DOI: 10.5220/0001670403870390

Multiplicity of Temporal Features. The most straightforward way of expressing temporal information is to relate it to calendar expressions. There are, however, other approaches that may be followed. For instance, some events my be related to some other events by means of temporal relations (,,A happens during B“).

Polymorphism of Time Expressions. Semantically equivalent temporal expression may be expressed in many different ways: ,,18 January 2007“, ,,18 I 2007” ,,18-01-2007“ ,,yesterday” (if reference date is 2007-01-18), relates to the same date.

Ambiguity/Imprecision. In many cases, there is no way to precisely qualify the value of temporal expressions, for instance, in fuzzy expressions, like ,,the beginning of May“.

2 RELATED WORK

The highlighted above issues are not fully addressed in the literature. The most well known system that deals with temporal IR is TOODOR (Temporal Object-Oriented Document Organization and Retrieval) (Llavori et al., 1998). Each article stored in TOODOR system is qualified by two attributes: publication date and temporal horizon (valid time), what makes it de facto bi-temporal database. Unfortunately, the temporal horizon is not defined. The authors state that its semantic is specific for particular application and its value should be set manually. In later publications (Llido; et al., 2001), it is suggested that its value should be based on calendar expressions extracted from document content. The indexing process consists of the following steps: extraction and normalization of all calendar expressions, determination of the most important date, definition of temporal horizon as an interval which covers all expressions that are within certain range from the most important date.

There is also a TDRM (Temporal Document Retrieval Model) (Kalczynski and Chou, 2005) which focuses mainly on fuzzy expressions, i.e. expressions whose value may not be determined precisely (e.g. ,,at the beginning of May“). The authors suggest using fuzzy set theory to encode their value. TDRM also accommodates Vector Space Model for weighting indexing terms. In this case each temporal reference is decomposed to a set of days. Each day is regarded as a single indexing term, whose weight is dependent on its frequency within the document and within the whole collection.

The major problems with the presented approaches are related mainly to the lack of: precise definition of temporal features, documentation of feature extraction and normalization process, explanation of rationale for undertaking certain design decisions (especially related to granularities conversion or terms weighting). Moreover, both approaches are constrained to a very limited set of temporal features.

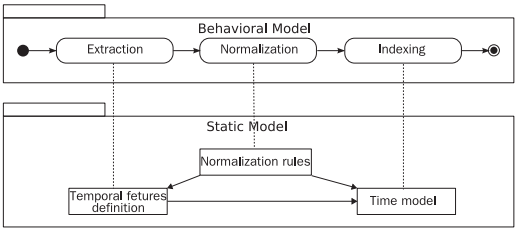

Figure 1: Metamodel of Temporal Information Indexing.

3 META-MODEL OF TEMPORAL INDEXING

Document index serves as a surrogate, which represents document important features in a compact and a machine processable form. The content of an index is dependent on the potential information needs. In most cases indexes cover: topics, words, or named entities important for an indexed document. Temporal index, on the other hand, should reflect time to which facts presented in the document relate.

Many potential and usable temporal indexing models exist. These models differ mainly in terms of: time model, definition of temporal features, their normalization and extraction procedures, definition of indexing terms, and definition of index structure. All these approaches may be described by one meta-model, which defines: necessary components, data they process, their interrelationships, and recommendations for certain design decisions (see figure 1). The presented meta-model is a result of generalization of existing approaches for temporal indexing and models that have been created during our experiments.

3.1 Static Model

The static model defines necessary resources required during indexing process, which include: a time model, a definition of temporal features along with extraction rules, and temporal features normalization rules.

A time model is the most fundamental component. It provides a definition of indexing terms, which may include: time points, intervals or granules. We suggest using a calendar-based time model. The decision

is motivated by its:

- Popularity – calendar-based temporal expressions occur relatively frequently, especially in news stories,
- Simplicity – one of the most common way of expressing temporal constraints by users is to use calendar expressions, using the same time model for queries and index simplifies the model,
- Expressiveness – model should allow to express semantics of temporal expressions as precisely as possible; each expression should be encoded at the granularity level at which it was expressed in a document.

A document may be then indexed with pairs (I,G), where I is a granule index within granularity G (see (Bettini et al., 1998) for calendar arithmetic). We suggest using following granularities: a day of the week, a day of the month, a week of the year, a month of the year, a quarter of the of year, a half of the year, a season of the year, a year, a decade, and a century – $G \in \{DOW \ldots MTH, YER \ldots CTR\}$. The choice is dictated by the relative frequency of expressions expressed at these granularity levels. The list obviously does not cover all potential granularities, for example: a day of the year and a fiscal year are missing, but they appeared relatively rarely in analyzed documents. The index I of granule within granularity G is computed as a number of granules between analyzed granule and reference granule. The reference granule for granularity days is the first day of this era. For other granularities, this is the granule that contains the day with index 1 ($DAY(1)$).

This construction has two advantages. Firstly, we do not lose semantics, whenautomatically shifting granularity levels (during ,,a week" is not the same as during six consecutive days that constitute this week). Secondly, it is easy to compare expressions on different granularity levels. For instance, in order to test if $MTH(i) \cap YER(j) \in \emptyset$, the process is trivial, while according to a definition of the calendar (Bettini et al., 1998) both MTH and YER are defined as a derivative granularities of granularity DAY.

The calendar is used to encode values of document temporal features. Following features have been defined:

Temporal Expressions. Temporal expressions relate directly to a model of time. All necessary information required to qualify their values is embodied in: the expression itself, the surrounding context, and the time model. No external knowledge is required. For instance ,,2007-01-02", ,,tomorrow" or ,,before" are temporal expressions, but ,,during Great Depression" is not one. Although, the last expression points to some time period, it requires knowledge at the beginning and ending dates of this event, in order to precisely set the time period.

Objects and Events. Objects and events posses temporal features. They themselves do not have a value specified by a time model but they exist in time. For instance, an event may have an occurrence date and an object exists during some time period.

Concepts. Concepts themselves, usually do not have a meaning allowing to relate them to certain time periods. We may assume, however, that conceptualization layer is dynamic. The new concepts are being created and some concepts lose popularity. Moreover, the popularity of the concepts appearing in documents change over time.

The last component used to characterize the indexing model is a normalization process. The normalization process sets values of temporal features in selected time model. In case of calendar model, for each temporal expression indices of granules and granularity level need to be specified. The normalization procedure is partially independent from the other components. It appears that more than one common normalization approach for different temporal features often exists, furthermore temporal feature may be normalized using different approaches. We can distinguish following normalization approaches:

Rules. For some categories of temporal features, it is possible to define normalization mechanism in terms of conditional statements (IF...THEN...rules). This approach is especially useful in case of calendar expressions. For example, if a reference date is ,,2000-01-01" and a date to be normalized is ,,February" and from thenarrative context it appears that we speak about future, then the year of the normalized date should be set to the year of the reference date, i.e. 2000.

DB of States/Events. Above, we have used an example of ,,Great Depression". The normalization of such an expression requires information at the beginning and ending dates of this event. It is possible to create a database of events/states, which may be in turn used for indexing purposes. The indexing model is certainly limited only to events/states it has knowledge on.

Distribution of Concepts in Time. We have assumed, that concepts used in text, or at least their subset, including concepts used to describe events and states are related to time. It is possible to build probabilistic model which defines

probability of occurrence of particular concept in documents related to different time points. One may use a joint probability to assess probability that a document containing certain concepts relates to a certain period.

3.2 Behavioral Model

The behavioral model defines which resources of the static model and in which order are to be used at each stage of the indexing process.

Generalizing investigated approaches, the indexing process consist of the following steps:

1. The processing unit is a single document. For each document a list of temporal features is extracted (according to the definition of temporal features). At this stage partial transformation or normalization of temporal features is possible. For instance, temporal expressions may be encoded using some formal notation. Therefore, the extraction process may be indirectly dependent on a time model.
2. Each of extracted features is normalized based on defined normalization process. The result of normalization is a list of temporal features values formalized with respect to the chosen time model.
3. Based on the normalized features temporal index is created. At this stage terms may be weighted or filtered.

4 SUMMARY

The metamodel denes the indexing process and resources that are necessary to accomplish it. Characteristic of a particular model are dependent on: a time model, a denition of temporal features and a normalization process. Having decided on calendar-based time model, we may look for promising models modifying denition of temporal features and normalization procedure. The following models were implemented with satisfactory results:

- Temporal references – We assume that an appearance of a temporal expression in text causes that the article is related to that date. We do not analyze, however, this relationship.
- Events – We assume that if an event occurs in a document, then the document itself is related to a date/dates specific for that event. Again, the semantic of this relationship is not analyzed. In this case a database of events and their specific dates is needed.
- Concepts – a probability of a concept occurring in a document depends on the period to which the document relates. In other words, documents that relate to different time periods may use diverse concept set. For instance, a concept ,,collective farming“ may occur with relative higher frequency in documents related to the rst half of he last century, then for example, in documents related to this century. Of course, one concept does not allow deriving any conclusions, but combining probability of occurrence of each concept contained in the document may give some clue on the document valid time.
- Semantic Similarity – In traditional IR systems indexing is sometimes based on the similarity of documents. In this approach, it is assumed that syntactically similar documents are also semantically similar and that semantically similar documents should have similar indexes. Therefore, syntactically similar documents should also have similar indexes.

REFERENCES

Bettini, C., Dyreson, C. E., Evans, W. S., and Snodgrass, R. T. (1998). A glossary of time granularity concepts. *Lecture Notes in Computer Science*, 1399.

Jensen, C. and Snodgrass, R. (2006). *Temporal Databases*.

Kalczynski, P. J. and Chou, A. (2005). Temporal document retrieval model for business news archives. *Inf. Process. Manage.*, 41(3):635–650.

Llavori, R. B., Cabo, M. J. A., and Barber, F. (1998). Discovering temporal relationships in databases of newspapers. In *IEA/AIE '98: Proceedings of the 11th International Conference on Industrial and Engineering Applications of Artificial In telligence and Expert Systems*, pages 36–45, London, UK. Springer-Verlag.

Llido;, D., Llavori, R. B., and Cabo, M. J. A. (2001). Extracting temporal references to assign document event-time periods. In *DEXA '01: Proceedings of the 12th International Conference on Database and Expert Systems Applications*, pages 62–71, London, UK. Springer-Verlag.