Qualitative and Quantitative Methods in Libraries (QQML) 4: 181-–189, 2015

Automated thesaurus population and management

E. Petraki, C. Kapetis, E. J. Yannakoudakis
Athens University of Economics and Business, Department of Informatics, Athens 10434, Greece

Abstract: FDB is a set theoretical model which allows the definition of multilingual databases and thesauri through a universal schema, offering administration utilities at both data and interface level, the definition of variable length objects, authority control, etc. The purpose of this paper is to investigate the issue of automatic thesaurus population of the FDB database as well as the automatic correlation between data records and thesaurus terms. The thesaurus forms part of the FDB model; more than one thesauri can be defined in FDB which can be multilingual or monolingual while the linking of each frame object (data record in terms of a traditional database) with the appropriate thesaurus terms can be achieved easily. In this paper we firstly describe briefly the FDB model, and proceed to present a) the algorithms which implement the linking of each frame object with the underlying thesaurus terms automatically, b) define algorithms for automatic thesaurus enrichment with terms derived from the data base, and c) outline research concepts and related work about the automatic thesaurus creation.
Keywords: databases, multilingual thesaurus, information retrieval, conceptual retrieval, FDB

1. Introduction

This paper suggests innovative algorithms and mechanisms which automatically associate and link each frame object (database record) in the FDB model with the related thesaurus terms of an underlying multilingual thesaurus. Moreover, the paper explains how an existing thesaurus of the FDB model can be enriched with new terms which are derived from the data frame objects. In addition, the paper discusses the problem of automatic thesaurus creation, research concepts and related work and proposes ways to automatically create a core thesaurus in the FDB model. Finally, the paper discusses future research.

This area of research is very important because it has to do with the linkage of frame objects in an FDB database with the appropriate thesaurus terms. The present research presents the algorithm that automatically establishes these links between frame objects and thesaurus terms by exploiting the structure of the

Received: 24.4.2014 / Accepted: 1.12.2014 ISSN 2241-1925

FDB model and the information provided by the thesaurus without requiring any knowledge from an expert. The proposed methodology completes all these correlations automatically while an expert may work manually to make additional correlations between data and thesaurus terms. This approach is very fast compared to the traditional linking methods which are carried out manually by experts. On the other hand, the proposed algorithm that automatically enriches a thesaurus with new terms derived from data frame objects can be used by specialists to add more terms to existing thesauri. An expert can also confirm that only the most relevant terms will be added in the thesaurus. In addition, the paper presents an algorithm that creates the basis for a new thesaurus. All the terms for the new thesaurus are derived from the underlying data frame objects of FDB and by using a dictionary. The relationship type "related term" is established between terms using the proposed algorithm.

Efficient management and exploitation of a thesaurus is very important in databases. In the FDB model it is possible to apply conceptual searches using the information provided by one or more multilingual thesauri. Current research offers an additional important tool to the FDB administrators and users

.

2. Related work – other research approaches

The issue of automatic thesaurus construction and enrichment has to do with a wide range of different research areas, including NLP-Natural Language Processing and Data Mining. NLP is a field of computer science, artificial intelligence and linguistics concerned with the interactions between computers and human (natural) languages (Wikipedia). Data Mining is a field at the intersection of computer science and statistics that attempts to discover patterns in large data sets; the analytic steps of knowledge discovery in Databases (Wikipedia). In NLP several different approaches are fields of research like natural language learning (NLL), natural language understanding, sentence breaking, word segmentation, word sense disambiguation, parsing (syntactic analysis), co-reference resolution, part of speech tagging (for those words that may have a different usage e.g. the word "book" as a verb and as a noun), relationship extraction etc. In what follows we present briefly some of the most important approaches in the field of automatic thesaurus construction.

AUTHECO is a research project aimed at developing information extraction for automatic construction of semantic networks and thesauri from corpus of domain specific texts, Panchenko A. (2008). This project investigates two problems of (semi-) automatic thesaurus generation from text corpora: the term extraction and the relationship extraction. Term extraction has to do with the selection of salient domain-specific terms from corpus and relationship extraction refers to the extraction of meaningful semantic relations between terms of a domain. It is based on two similarity measures Serelex and PatternSim and it is a kind of "lexico-semantic search engine"; given a text it provides a list of related words.

Panchenko, Adeykin, P. Romanov and A. Romanov (2012) presented methods of extraction of semantic relations between words relying on the k-nearest algorithms and two similarity measures to extract relations from the

abstracts of Wikipedia articles. The input data for these methods is a set of definitions for each input concept and the output is a set of terms. The main contribution and the implementation of this research is an open source system which effectively calculates semantic similarities between the input concepts and builds a list of semantic relations. This method can be used to extract relations between concepts in English or other languages in Wikipedia but uses language dependant resources such as stop-lists, part-of-speech tagger and lemmatizer.

Panchenko, and Morozova (2012) also described and evaluated several novel hybrid similarity measures in the context of semantic relation extraction and the results have shown that the hybrid methods outperform the single measures. Grefenstette (1993) presented a number of automatic techniques that produce a first draft of a thesaurus from any domain-defining collection of text. The techniques are knowledge-poor in that no domain knowledge is required for their use. Some researchers use hybrid measures, which do not return the types of relation between terms but they assume the retrieval of synonyms, hypernyms, co-hypernyms etc. Some other approaches produce reliable results but they must be corrected and checked by an expert.

Our approach introduces algorithms which correlate frame objects of each FDB database with the appropriate terms of a thesaurus in the FDB model, enrich the existing thesaurus with terms derived from the database and create a base of related terms which can form the core for a new thesaurus.

Finally, section 3 presents briefly the FDB model, section 4 illustrates the algorithms for thesaurus population and management and section 5 outlines the advantages of the proposed approach and the issues for future research.

3. The FDB Model

FDB is an integrated set-theoretic model for database systems that forms a framework for defining a structure that eliminates completely the need for reorganization at the logical level, E. J. Yannakoudakis, P. K. Andrikopoulos (2007) and Yannakoudakis E.J., Tsionos C.X. and Kapetis C.A (1999). FDB provides a universal model which allows the definition of any database by specifying the appropriate metadata without requiring the change of the underlying schema. Amongst other utilities, FDB allows administration of multilingual databases at both data and interface levels, definition of variable length objects (records in the traditional sense), etc. Any changes that may be necessary at the data level do not affect the universal database schema but simply the identification of the appropriate metadata. The basis for the creation of the unified schema is the definition and manipulation of metadata that compose the whole structure, E. J. Yannakoudakis (1987). Accessing the information from an FDB schema becomes very easy with the use of simple statements provided by the Conceptual Universal Database Language (CUDL), Yannakoudakis E. J., and Nitsiou M. (2006). The FDB universal schema can be used to define one or more multilingual thesauri and provides, besides traditional keyword search, conceptual searches through one or more multilingual thesauri, E. Petraki, C. Kapetis, E. J. Yannakoudakis (2013).

The basic elements of the model are based on the mathematical theory of unordered sets and consist of the following sets: a) entities: the unordered set of registered entities that participate in the logical schema, b) tags: the set of attributes describing each entity, c) subtags: the set of simple atomic attributes which constitute existing complex tags, d) domains: the set of all data domains, e) languages, vocabulary, messages: sets of strings or coded values that present human languages and corresponding messages, E. J. Yannakoudakis (1987). (See reference [1] for a short example).

4. Algorithms for thesaurus population and management

In what follows we present three basic algorithms for the management and the population of thesaurus. The first one links each frame object of the database (see the bibliographic database example in reference [1]) with the appropriate thesaurus terms. The second one can be used to enrich the entire thesaurus with new terms (e.g. synonyms). These terms are derived from specific database tags – in the example the tags that may contain important terms are the abstract, the title and the keywords of a book or article (tags 601, 602 and 603). The third algorithm creates a core of terms for a new thesaurus from the frame objects of the FDB database.

Input for the algorithms: a) The *frame_entity_numbers* and the *tag* numbers of all important entities and corresponding tags for the **data**, the **thesaurus**, the **relations** between terms and the linking between data and thesaurus terms. The frame_object_number and its appropriate tags of the FDB database are used for indexing or thesaurus enrichment-construction. In the example of reference [1], a bibliographic database is stored in the FDB model. The tags which store the title, the abstract and the keywords of an article or a book are used as input data for the algorithms. These tags are used to relate each frame object (data record) with the appropriate thesaurus terms. Furthermore these tags can be used to provide new terms in order to enrich any existing thesauri. The algorithms presented below accept the following input:

Frame entity numbers: **DE** (Data Entity) = 100 (book/article), **TE** (Thesaurus Entity) = 1 (thesaurus terms), **RE** (Relation Entity) = 2 (describes the relationship types between thesaurus terms), **DFO** = Data Frame Object, **TFO** = Thesaurus Terms Frame Object, **RON** = Relation Object Number of the "related term" relationship type.

Tags: **DTAGS** (Data Tags) = 601 (abstract), 602 (title) and 603 (keywords) tags of the 100 (book/article) entity, **LTAG** (Linking Tag) = 650 of the 100 entity. LTAG is used to link each data frame object with the appropriate thesaurus term(s). **TTAG** (Thesaurus Term Tag) = 200 (term tag) of the 1 (thesaurus terms) entity.

b) Stopwords: a list of stop words for the English language, like http://norm.al/2009/04/14/list-of-english-stop-words/ or http://www.ranks.nl/resources/stopwords.html. Stopwords are considered the articles or other words which have an ambiguous meaning, therefore they can't be used to correlate data with thesaurus terms.

c) Dictionary: Since the base language is English, we need an English dictionary like http://dictionary.reference.com/ which provides synonyms and other related words except for an extended explanation of each word.

A) Correlating data frame objects with the thesaurus terms

Algorithm A

Define **T**, including all thesaurus terms (all terms in tag 200)

For each DFO

F = ‘’ // **F** includes the text of all specified tags

For each tag in DTAGS

F = F +’ ‘+tag

End loop

W = {} // Define the set of the words **W**

For each tag in DTAGS

W = W ∪ tag

End loop

For each w ∈ W *// Remove stopwords*

If w ∈ S then

W = W – {w}

End if

End loop

For each word w in W

If w ∈ T then

Link_Tag*(DE, TE, DFO, LTag, TFO)*

else if w is part of t ∈ T then

if t in F then

Link_Tag*(DE, TE, DFO, LTag, TFO)*

end if

else

Find all synonyms in Dictionary(w,SYN)

For each syn∈ SYN

If syn ∈ T then

Link_Tag*(DE, TE, DFO, LTag, TFO)*

Add_Term*(TE, TTAG ,TFO, w, RE, RON)*

exit loop

End if

End loop

End If

End Loop

End A

Link_Tag(DE, TE, DFO, LTag, TFO): uses the linking tag (LTAG, 650 in the example) to correlate the current frame object with a specific thesaurus term. Sets TFO in the LTAG of the current DFO.

Add_Term(TE, TTAG, TFO, w, RE, RON): adds the word *w* in TERM_DATA and in THESAURUS_TERM_DATA as a related term

Explanation: Algorithm A searches each word *w* derived from the specified data tags in the thesaurus terms *T*. If the word *w* is a thesaurus term then the algorithm adds the term's object number into the appropriate data tag (LTag) of the specific frame object. If the word *w* does not exist in the thesaurus then the algorithm checks if the *w* is part of a specific thesaurus term *t*. In this case, if the term *t* exists in *F* (the text of all the data tags), the algorithm correlates the current frame object with this term *t*. Otherwise, the algorithm searches in a dictionary to find synonyms of the word *w*. If a synonym exists in the thesaurus, then it correlates the current frame object with the current term and adds the *w* into the thesaurus as a related term for the existing synonym.

B) Enrichment of an existing thesaurus

Algorithm B

Define **T**, including all thesaurus terms (all terms in tag 200)

For each DFO

W = {} // Define the set of the words **W**

For each tag in DTAGS

W = W ∪ tag

End loop

For each w ∈ W *// Remove stopwords*

If w ∈ S then

W = W – {w}

End if

End loop

For each word w in W

Find all synonyms in Dictionary(w,SYN)

If wϵT then

For each syn ∈ SYN

If syn ∉ T

Add_Term*(TE, TTAG ,TFO, syn, RE, RON)*

End if

End Loop

else

For each syn ∈ SYN

If syn ∈ T

Add_Term*(TE, TTAG ,TFO, w, RE, RON)*

exit loop

End if

End Loop

End if

End loop

End loop

Algorithm B searches each word *w* derived from the specified data tags in the thesaurus terms. If this word *w* exists in the thesaurus then the algorithm adds every synonym of *w* into the TERM_DATA and

THESAURUS_TERM_RELATIONS under the same frame_entity_number of the existing thesaurus term and defines this synonym as a "related term". If *w* does not exist in thesaurus then the algorithm searches all the synonyms of *w* in the dictionary. If a synonym exists in thesaurus then the algorithm adds this word *w* into TERM_DATA and THESAURUS_TERM_RELATIONS as a "related term" of existing synonyms.

C) Creating a new thesaurus from FDB database records

The issue of automatic thesaurus construction forms a very important research area. In this section we present the process for creating the core of a thesaurus for a specific domain by using the data stored in the FDB database. This algorithm represents the initial stage for creating the thesaurus automatically, although more research is required in this area. The results of Algorithm C form the basis for a new thesaurus which can be extended and integrated by an expert.

Algorithm C
For each DFO
W = {} // Define the set of the words **W**
For each tag in DTAGS
W = W ∪ tag
End loop
For each w ∈ W *// Remove stopwords*
If w ∈ S then
W = W – {w}
End if
End loop
For each word w in W
Add_Term*(TE, TTAG ,TFO, w, null, null)*
Find all synonyms in Dictionary(w,SYN)
For each syn ∈ SYN
Add_Term*(TE, TTAG ,TFO, syn, RE, RON)*
End loop
End Loop
End C

This algorithm adds each word *w* derived from the specified tags of the data frame objects into the thesaurus terms of a new thesaurus entity. Then, it adds the synonyms of *w* into the new thesaurus and defines them as "related terms". All metadata necessary for the creation of a thesaurus (thesaurus entity, thesaurus relation types) must be defined (or copied from an existing thesaurus) by the administrator.

5. Advantages and issues for further research

FDB is an integrated database management system which provides a universal schema that allows the definition of any multilingual database and

thesaurus by setting the appropriate metadata, while allowing data retrieval by using both free text techniques and conceptual searches through the use of multilingual thesauri.

In a traditional database system, the correlation between the data records and the thesaurus terms requires a great deal of effort, time and expertise. The research presented here facilitates the creation and population of a thesaurus, which can be implemented in any FDB database. This approach is very useful especially for large bibliographic databases as it provides the automatic correlation of data and thesaurus terms without requiring any effort on behalf of the administrator. Another advantage of the proposed method is that the user can run the same algorithm for different thesauri and dictionaries. This provides significant flexibility because each data frame object can be assigned to many thesaurus terms from different multilingual thesauri. The second algorithm (algorithm B) gives an additional tool to the DBA because it enables the enrichment of existing thesauri with new related terms. All new terms are derived automatically from the specific tags of the FDB database as proposed by an expert. The third algorithm (algorithm C) provides all related words derived from an FDB database which can form the core for a new thesaurus of a specific domain.

The algorithms presented here can be expanded and improved by including language dependant rules for words derived from FDB databases. A future version of these algorithms can contain mechanisms to recognize the plural form and the root of words or apply syntactic analyses in order to find the most significant words in a sentence. With such rules it will be possible for the algorithms described above to correlate more frame objects with thesaurus terms and enrich an existing thesaurus with even more new terms.

References

E. Petraki, C. Kapetis, E.J. Yannakoudakis, Conceptual database retrieval through multilingual thesauri, Computer Science and Information Technology 1(1): 19-32, 2013

E. J. Yannakoudakis, P. K. Andrikopoulos, A set-theoretic data model for evolving database environments, In Proceedings of the International Conference on Information & Knowledge Engineering, IKE 2007, June 25-28,2007 Las Vegas, Nevada, USA.

Yannakoudakis E.J., Tsionos C.X. and Kapetis C.A, A new framework for dynamically evolving database environments, Journal of Documentation, Vol. 55, No. 2, pp. 144-158, 1999.

E. J. Yannakoudakis, An efficient file structure for specialised dictionaries and other 'lumpy' data, International Journal of Information Processing & Management, Vol. 23, No. 6, pp. 563-571, 1987.

Panchenko A., Adeykin S., Romanov P., Romanov A., "Extraction of Semantic Relations between Concepts with KNN Algorithms on Wikipedia". In Proceedings of Concept Discovery in Unstructured Data Workshop (CDUD) of International Conference On Formal Concept Analysis, pp.78-88, Belgium, 2012

Panchenko A., Olga Morozova**,** A study of hybrid similarity measures for semantic relation extraction in Proceeding of HYBRID 2012 Proceedings of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data, Pages 10-18, Association for Computational Linguistics Stroudsburg, PA, USA

Grefenstette G., Automatic thesaurus generation from raw text using knowledge-poor techniques. In making sense of words, ninth annual conference of the UW centre for the OED and text research, 1993

Wikipedia http://en.wikipedia.org/wiki/Natural_language_processing

Wikipedia http://en.wikipedia.org/wiki/Data_mining

Panchenko A., AUTHECO: AUtomatic THEsaurus COnstruction, a research project supported by Université Catholique de Louvain and Wallonia-Brussels International (WBI), http://cental.fltr.ucl.ac.be/projects/autheco/, October 2008 (start).

Yannakoudakis E. J., and Nitsiou M., A new conceptual universal database language (CUDL), In 2nd International Conference From Scientific Computing to Computational Engineering, Athens, Greece , 2006