Knowledge Discovery from Texts: A Concept Frame Graph Approach

Kanagasabai Rajaraman*
Laboratories for Information Technology
21 Heng Mui Keng Terrace
Singapore 119613
kanagasa@lit.a-star.edu.sg

Ah-Hwee Tan
Laboratories for Information Technology
21 Heng Mui Keng Terrace
Singapore 119613
ahhwee@lit.a-star.edu.sg

ABSTRACT

We address the text content mining problem through a concept based framework by constructing a conceptual knowledge base and discovering knowledge therefrom. Defining a novel representation called the Concept Frame Graph (CFG), we propose a learning algorithm for constructing a CFG knowledge base from text documents. An interactive concept map visualization technique is presented for user-guided knowledge discovery from the knowledge base. Through experimental studies on real life documents, we observe that the proposed approach is promising for mining deeper knowledge.

Categories and Subject Descriptors

I.2.4 [**Artificial Intelligence**]: Knowledge Representation Formalisms and Methods; I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Text Analysis*

General Terms

Algorithms,Experimentation

Keywords

Text mining, knowledge extraction, concept mapping, information visualization.

1. INTRODUCTION

Text Mining or Knowledge Discovery from Texts[3] is an area that is gaining importance due to its potential in offering tools to effectively tackle the information overloading problem. Text Mining methods can be broadly classified as solutions for two tasks namely, document browsing and content mining. Browsing methods usually work at the document level employing coarse-grained mining techniques whereas content mining works at a deeper content level analyzing syntactic and semantic aspects of the text. In this paper, we propose a new text content mining algorithm using efficient text processing techniques and develop a knowledge discovery system based on this algorithm. Our approach is through the following two step process: *Conceptual Knowledge Base(CKB) Construction* in which a structured database of important concepts and relationships is extracted from the texts, and *CKB Mining* where mining techniques are applied on the knowledge base to discover useful information.

*Corresponding author

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
CIKM'02, November 4–9, 2002, McLean, Virginia, USA.
Copyright 2002 ACM 1-58113-492-4/02/0011 ...$5.00.

2. CONCEPT FRAME REPRESENTATION

The conceptual knowledge base is represented through *Concept Frames*, defined below.

DEFINITION 2.1. *A concept frame is an object (*NAME, SYNSET, RELS, CONTEXTS*) where:* NAME *is the name of the concept,* SYNSET *is a set of synonyms of the concept,* RELS *is a set that describes the relations of this concept with other concepts through relation tuple of the form (AgentCF, rel, ObjectCF), where 'AgentCF' and 'ObjectCF' are pointers to concept frames and 'rel' is a relation between them, and* CONTEXTS *is an (optional) set of text segments corresponding to each relation tuple in* RELS.

Through the RELS field, a set of relationships of a concept frame with other concept frames is defined. Hence, a concept frame can be thought of as a node of a graph in which the edges are defined by RELS. We call this graph a concept frame graph, formally defined below:

DEFINITION 2.2. *Given a set of concept frames* $F_i = (Name_i, Synset_i, Rels_i, Contexts_i)$, *with* $Rels_i = \{(a_{ij}, r_{ij}, o_{ij}), j = 1, \cdots, M_i\}$, $i = 1, \cdots, N$, *the Concept Frame Graph(CFG) is a finite, directed, edge-labeled graph in which* F_i *and every* a_{ij}, o_{ij} *other than 'Self', define the nodes and the pairs* (a_{ij}, o_{ij}) *define the corresponding edges with* r_{ij} *as the edge label.*

3. CKB CONSTRUCTION

Given a document collection, the set of concept frames that completely describes this collection would form a CFG, according to Definition 2.2. We use this CFG as the Conceptual Knowledge Base(CKB) of the document collection.

Below we describe an algorithm for learning this CFG from a set of documents, through a sequence of steps.

Pre-processing: The documents are first pre-processed to remove menu bars (as found on web pages) or formatting specifications such as HTML tags so that only the main body of the content is retained.

Name Entity (NE) Recognition: This step extracts all entities such as person names and company names. Different variations of an entity are identified through a co-reference resolution algorithm[7]. The variations are reduced to a standard form and the documents annotated using this standard form for further processing.

NVN 3-tuples extraction: An NVN 3-tuple is of the form (NC, VC, NC), where NC is a Noun Clause and VC is a Verb Clause. NC and VC are extended forms of Noun Phrase and Verb Phrase respectively, defined by the regular expressions

NC :- $(ADJP)?(NP)^{+}((IN|VBG)^{*}(ADJP|NP))^{*}$

VC :- $(ADVP)?(VP)^{+}(IN|ADVP)^{*}(VP)^{*}$

where ADJP is an adjective, NP is a noun phrase, IN is a preposition, VBG is a Verb gerund, ADVP is an adverb and VP is a verb phrase. The NVN 3-tuples will be ultimately used to generate the Synset and Rels parts of the concept frames.

As a first step in the extraction, text is tagged using an in-house developed Part-of-Speech (POS) tagger. Then we employ a handcrafted rulebase to extract the NVN 3-tuples. Then, the NC's are sense disambiguated and then the synsets learned through a clustering algorithm, as described below.

Sense Disambiguation: Our disambiguation algorithm is based on WordNet[5]. For every word, WordNet distinguishes between its different word senses by providing separate synsets and associating a sense with each synset. Our algorithm makes use of the context of the words in an NC to compute a distance measure and pick the correct word sense.

Clustering: In this step, we group the disambiguated NC's through a clustering algorithm. We employ a fuzzy ART[2] based clustering algorithm. A fuzzy ART system includes an input field that represents a current input vector; a field F_1 that receives both bottom-up input from the input field and top-down input from a field, F_2, that represents the active code or category. A detailed description of the ART network and the learning algorithm can be found in [2].

Fuzzy ART formulates recognition categories of input patterns by encoding each input pattern into a category node in an unsupervised manner. Thus each category node in F_2 field encodes a cluster of patterns. Hence, the clustering problem translates to the problem of the creation of new categories in the F_2 field as more patterns are presented. In our problem, the input patterns correspond to the NC's. Each NC is converted to a vector form for processing by ART. All key terms (omitting stop words) are extracted from the part-of-speech information to form a weight vector, $\mathbf{c} = (c_1, c_2, \ldots, c_M)$, where M denotes the number of features extracted and c_i denotes the term frequency for term $i, i = 1, \cdots, M$. The vector is then normalized by dividing all elements with $max_i\{c_i\}$.

3.1 Frame Filling

Each cluster generated in the previous section, is mapped to a concept frame through a Frame Filling algorithm. In this algorithm, the cluster members are first collected to form the synset. Then, by identifying the synset to which the agent and object NC's of an NVN tuple belong to and replacing them with the corresponding cluster ID's, the NVN 3-tuples are generalized to form the Rels. The Contexts field is formed by collecting the sentence fragments corresponding to each relation tuple. A key subphrase is extracted from the most dominant member of the synset as the name of the frame.

The above procedure results in a set of concept frames being extracted from the given document collection. By Definition 2.2, this constructs the Concept Frame Graph (CFG) and hence the Conceptual Knowledge Base.

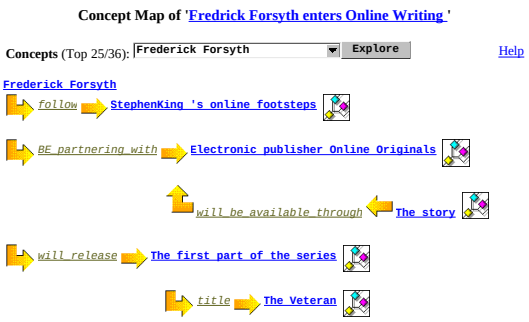

Figure 1: Concept Map Visualization.

4. CKB MINING

We present a concept map visualization interface for user-guided knowledge discovery. The interface consists of two parts: a pull down menu and a concept map (See Figure 1). The menu displays the top 25 dominant concepts ranked using synset cardinality. With the most dominant concept as the root, a horizontal tree of all concepts connected to it, together with the relations (as edge labels), is drawn. Since the tree could be potentially large, the tree is drawn only up to depth 2, but hyperlinks to explore the map further are provided. Each concept name can be clicked to view it's synset. Similarly, a relation can also be clicked to display the context of this relation (from the Contexts field). An additional link is provided to display the complete concept frame.

The concept map provides a user-friendly interface for viewing the document content. As illustrated above, the user can choose a concept and view its relationship with other concepts and thus a theme-based view of the documents is possible. In fact, this implies that our interface permits different perspectives of the content. This we believe is an important feature of our method that facilitates user-guided knowledge discovery.

5. EXPERIMENTAL STUDIES

For the experiments, we have used technology news articles collected from CNet (http://news.cnet.com) and ZDNet (http://www.zdnet.com/zdnn/). In particular, document sets on three topics have been considered, namely Goner Worm (3 documents), Nimda Worm (5 documents), and Napster in Court (10 documents).

We consider a Q & A task in which the ability of our system in enabling users to find answers effectively is evaluated.

Doc Set	Avg Total Score		% Improvement
	DBS	Concept Map	
Case I	0.549	0.733	+33.5 %
Case II	0.599	0.833	+39.0 %
Case III	0.683	0.716	+4.8 %

Table 1: Summary of Results on Q&A task.

A *domain expert* sets questions (that could be answered by human reading of the documents) to be answered by the three *experimental users*.

Given a document set, the three users were shown the concept map generated and asked to answer the five questions. They were also asked to mention how easy it was to find the answers in terms of 5 levels: *Very easy, Rather easy, Rather Tough, Very Tough, Can't find answer*. To arrive at a quantitative measure, we score these 5 levels respectively as: *1.0, 0.75, 0.50,0.25 and 0.0*. The score is halved for a partially correct answer (decided by the expert) and a wrong answer is scored 0.0. For each question, an average score from the responses of three users is computed to measure the overall effectiveness of the system for this question. For baseline comparison, the candidates were asked to repeat the answering task using a simple Document Browsing system (DBS) in which the user can browse the title, summary or the full content of the document to find the answers. The results are summarized in Table 1 by averaging the scores over the five questions under each case.

Both the methods performed near par in Case III where the questions were easier. However, in case I and II containing relatively tougher questions, we observe that the concept map has enabled the answers be found more efficiently. This is made possible by our concept map based visualization of the documents. Given a question, the user can observe which concept the question is talking about, and can relate to a concept shown on the map interface. Then the user can quickly focus on that concept by clicking and exploring it, skipping unnecessary parts of the text.

A more detailed account of our studies is available at http://textmining.lit.org.sg/people/kanagasa/cfg/.

6. RELATED WORK

R. Feldman, *et.al.*[4] have proposed a system for extracting terms and computing associations between them to display a *Context Graph*. The context graph can be used to discover interesting term relationships. The context graph primarily uses numerical strengths to capture the relations. ClearResearch (http://www.clearforest.com) is an improved version of their system that makes use of an adhoc rulebase to support specific named relations in context graphs. Byrd and Ravin[1], in a similar vein, have implemented a system for extracting a network on concepts and relations. The 'concepts' have been defined informally and are usually entities or terms. (Thus, the *triples* defined in their paper are similar to the NVN 3-tuples, which are low level text extracts.) Semio Map (http://www.semio.com) and TextAnalyst (http://www.megaputer.com) are some of the other systems that can create graphs based on term correlations. In contrast, our approach focuses on extracting richer named relations in a more formal framework. Moreover, we have followed a conceptual approach that is based on generalisation of terms through synsets and sense disambiguation.

Liddy, *et.al.*[6] define a representation called the Concept-Relation-Concept (CRC) triples which are similar to our relation 3-tuples (See Definition 2.1). The system analyzes raw text to construct a database of CRC triples which can be used for, e.g. querying tasks. Though this system appears to closely resemble our work, their concept identification system requires a Conceptual Hierarchy Database constructed a priori. Our system does not require constructing any such domain dependent databases and we can identify concepts directly from the input text documents. In addition, the relations in CRC triples are semantic relations whereas concept frames capture syntactic relations. Semantic relations, though intuitively more appealing, are usually difficult to extract and may be computationally expensive.

7. CONCLUSION

In this paper we have addressed the problem of text content mining through a concept-based framework. We have proposed a new concept representation called the Concept Frame Graphs and presented a novel learning algorithm for constructing CFG's from texts. A user-friendly visualization interface has been developed for interactively mining the CFG. The utility of our approach has been illustrated through experimental studies on real life documents. The results are only preliminary as the number of documents considered is rather small. Currently we are evaluating our system on bigger collections with more candidate users. We are also exploring several ways to extend our work such as building richer and more accurate knowledge bases, mining more sophisticated relations than lexico-syntactic and implementing 2D and 3D JAVA based visualization interfaces.

8. ACKNOWLEDGEMENTS

The authors would like to thank Jian Su and Guo-Dong Zhou for the Multilingual Efficient Analyzer(MEA) API, which was made use of in NE Recognition and POS-tagging.

9. REFERENCES

[1] R. Byrd and Y. Ravin. Identifying and extracting relations from text. In *VLDB '99*, 1999.

[2] G. A. Carpenter, G. Grossberg, and D. B. Rosen. Fuzzy ART: Fast stable learning and categorization of analog patterns by an adaptive resonance system. *Neural Networks*, 4:759–771, 1991.

[3] R. Feldman and I. Dagan. Knowledge discovery in textual databases (KDT). In *Proceedings of KDD-95*, pages 112–117, 1995.

[4] R. Feldman, M. Fresko, H. Hirsh, Y. Aumann, O. Liphstat, Y. Schler, and M. Rajman. Knowledge management: A text mining approach. In *Proc. of the 2nd Int. Conf. on Practical Aspects of Knowledge Management*, Basel, Switzerland, 1998.

[5] C. Fellbaum. *WordNet: An electronic lexical database*. The MIT Press, Cambridge, Massachusetts, 1998.

[6] W. Paik, E. D. Liddy, J. H. Liddy, I. H. Niles, and E. E. Allen. Information Extraction System and Method Using Concept-Relation-Concept (CRC) Triples. US Patent 6,263,335, Jul, 2001.

[7] G. D. Zhou and J. Su. Named entity recognition using a HMM-based chunk tagger. ACL '02, 2002.