Applying Formal Concept Analysis to Teaching Material Extraction

Shao-Chun Li
Department of Information Communication, MingDao University, Taiwan
learry@gmain.com

Ko-Kang Chu
Department of Information Communication, MingDao University, Taiwan
kirk@ms2.hinet.net

Maiga Chang
School of Computing & Information Systems, Athabasca University, Canada
maiga@ms2.hinet.net

Abstract: Text summarization system can save the time for user when reading large number of documents. The summary of text summarization system usually composed of meaningful sentence which represent content of text. The relations between keyword usually come from their co-occurrences in document. This study using hierarchical clustering method cluster sentences and apply concept formal analysis to find out the implications between keywords. The position of sentence appears in document also influence the importance of sentence. Finally the system selects sentences which represent document according to the weight of keywords, implications between keywords and position in document. In this research, we present an automatic text summarization system which can extract important keywords from document automatic and offer a short summary represent document.

Introduction

The principle of text mining is deriving useful knowledge or high quality information from unstructured or semi-structured text. Several techniques have been proposed for text mining such as natural language processing (NPL), conceptual structure, association rule mining, and information extraction. The results of text mining can be such as knowledge, interesting rules, and summary of document.

The main principle of Information Extraction (IE) system is obtained or recognized useful information, relationships or events form document automatically. The result of information extraction can be interesting rules, interesting items or well-defined data from specific domain. After structured data obtained, the information extraction system represented information in clear and united form and can be use for further process.

In general, information extraction usually involves the process of structuring the input text, deriving pattern within structured data, and finally representing output. Cardie (Cardie, 1997) defined five progresses of information extraction: tokenization and tagging, sentence analysis, extraction, merging and template generation. Each input text has been divided into words and disambiguated or tagged each word with respected to part-of-speech class. The sentence analysis phase is grouping words into nouns group, verbs group, pronouns groups, adjectives groups, adverbs groups and conjunctives groups. In the extraction phase, system identifies the relations among relevant entities in the text and merges relations in the merging phase. Finally, the relations interpreted in clear and understanding form to user.

Automatic summarization of articles has two main objectives: the first objective is how a summarizer handles huge data; and, the second objective is how the system produces human readable summary (Amini, Usunier, and Gallinari, 2005). There are two major summarizer types: the abstracts and the extracts (Havy and Lin, 1999). The abstract type summarizer needs to understand document and language generation in order to generate summary, the process is complex and hard to apply to on-line document (Sparck-Jones, 1993). The extract type summarizer only uses the portion of original document with simple analysis process. There are many different extract type summarizers use different analysis methods, such as keyword-based (included topics), query-based (user-specific), and document structure-based methods. The keyword-based summarizers focus on the frequency of keywords, relations among keywords, and keyword locations on the document. The query-based summarizers focus on the interactions between users (Daumë III, 2006; Varadarajan and Hristidis, 2006). There are many technologies can apply to text summarization such like information retrieval, query expansion, latent semantic analysis, neural network, fuzzy theory, and natural language processing, these technologies are extracting summary in semantic level.

This research presents an information retrieval-based text summarization system which considers about keyword expansion, phases and document structures. The system doesn't need users to input keywords but generate keywords with analyzing the sentences on the document. Formal concept analysis is applying to this system in order to find out the implications which contained in the sentences. The implications represent subordinate relations between keywords and can be used for keyword expansion. At last, the document structure is used to adjust sentence weights. There are some parameters used to expand keywords and compose the summary structure.

Related Works

A text summarization system has to represent the document contents with short sentences no matter it uses abstract or extract technology. Most of summary systems are using information retrieval technology and have four problems needed to solve: phrase, polysemy, synonymy, and term dependency. Part-of-Speech tagger can analyze each sentence and give each word a tag according to how the word is used in sentence. Formal concept analysis can figure out the implications between keywords and show the term dependencies out. This research uses these three technologies to solve the four problems and develops an automatic text summarization system.

The summary is the highest quality text of a document and representing the content of a document. There are two major types of summaries: abstracts and extracts. Extracts are summaries composed of sentences copied form the input document (e.g.,15% of original document). Abstracts are summaries composed of sentences not present in the input document.

Automated summarization techniques tried during the 1950's and 1960 was started with pioneering work by Luhn (Luhn, 1958). Luhn purposed a statistical approach to text summarization based on term (i.e. keyword) frequency and term normalization. Since then, many approaches have been purposed. Edmundson (Edmundson, 1968) purpose three addition features to improve the keyword features purposed by Luhn: *Lexical cues*, *Position in text*, *Sentence location*.

Although each of these approaches provides some utility or strategy for text summarization but they all depend on the format and style of writing. These approaches usually work in newspaper, magazine article, or formatted documents and do consider the relevant between keywords, sentences, etc.

Formal Concept Analysis (FCA) was proposed by Rudolf Wille in 1982 and widely used in many different domains such like psychology, sociology, biology, mathematics, industrial engineering, and computer sciences (Wolff, 1994). Formal Concept Analysis is a method which identifies conceptual structures and produces graphical visualizations of the inherent structures among data sets. The mathematical lattices (also called concept lattices) are used in Formal Concept Analysis which can be interpreted as classification system.

In Formal Concept Analysis, the data set is composed by formal context K=(G, M, I). *G* is the object set, *M* is the attribute set and *I* is the binary relation between object and attributes. An example of concept context is shown in tab. 1.

	Flying	Feather	Animal	Quadruped	Wings
Bat	1	0	1	0	1
Cat	0	0	1	1	0
Dog	0	0	1	1	0
Eagle	1	1	1	0	1
Ostrich	0	1	1	0	1

Table 1 Example of Concept Context

In this case, the object set *G*={Bat, Bird, Cat, Dog, Ostrich} and the object set *M*={Fly, Feather, Animal, Quadruped, Wings}, and the binary relations *I* is shown as a matrix which 1 represents objects g have attribute *m*. In formal context, G and M has following definition: (1) if $A \subseteq G$ then $A' = \{m \in M \mid \forall g \in A, (g,m) \in I\}$. Each elements in object set A all have corresponding attributes in set A'. For example, $\{\text{Bat}, \text{Eagle}\}' = \{\text{Animal}, \text{Fly}\}$. (2) if $B \subseteq M$ then $B' = \{g \in G \mid \forall m \in B, (g,m) \in I\}$. Each elements in attribute set B all have corresponding object in set B'. For example, $\{\text{Feather}, \text{Animal}\}' = \{\text{Eagle}, \text{Ostrich}\}$.

The formal concept is a pair (A, B) which has following relations: (1) $A \in G$ and $B \subseteq M$ (2) $A' = B$ and $B' = A$. If a pair set (A, B) satisfied the above formula we can say pair (A, B) is a formal concept; A is extent and B is intent. Take table 1 as example, there is a formal concept pair$\{\{\text{Cat}, \text{Dog}\}, \{\text{Animal}, \text{Quadruped}\}\}$. Cat and dog have intention attributes Animal and Quadruped; the extension of attributes Animal and Quadruped is Cat and Dog.

The formal concepts can derived from concept context and used to build lattice. The concept lattice has following definition: a set of all formal concepts of a formal context K with the partial order $(A_1, B_1) \le (A_2, B_2) \Leftrightarrow A_1 \subseteq A_2$ and $(A_1, B_1) \le (A_2, B_2) \Leftrightarrow B_1 \supseteq B_2$. Fig. 1 is an example of concept lattice from concept context in table 2.3. There are several implications in the concept lattice. For example each object have attribute "Flying" also have attribute "Wings".

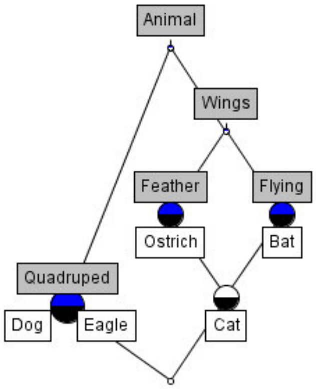

Figure 1 Example of Concept Lattice

Text Summarization

Keyword is the key to do text summarization. The first step to summarize is to find useful keywords. Words have many tenses and meanings in sentences; the first step of keyword selection is removing the tense and non-meaningful words then finding meaningful phases out. Secondly, selecting important keywords according to frequencies occurred in text. The keyword associations and sentence positions in original text also influence the importance of sentences. Before calculating the sentence weights, the sentence keywords are expanded according to its associations and the sentence length. Finally, the sentence weights are adjusted according to the sentence positions in the text.

Keywords are those words contain most important concepts of the text (or the document), however, not all words can be selected as keywords. The word importance will increase with the frequency in the text. An important thing is that the words with highest frequency are not keyword; keywords are those words with high frequency and meaningful. Therefore, the first step of word selection is removing those words with highest frequency and meaningless.

This study the confidence function *conf*() is used to decides whether to merge two words or not. Confidence is conditions probability method and usually used to explain the reliability of an association rule or item set from transactions in data mining. If we think document as database and sentences as transactions then we can use confidence to find the reliability of new phase. The formula of *conf*() isshown as bellow.

$$conf(t_i t_j) = \max(P(t_i t_j \mid t_i), P(t_i t_j \mid t_j)) = \max(\frac{|s_{t_i t_j}|}{|s_{t_i}|}, \frac{|s_{t_i t_j}|}{|s_{t_j}|}) \qquad (1)$$

$\forall\, t_i \in nouns, verbs, adjectives, adverbs, ...,, \ \forall\, t_j \in nouns$

$|s_{t_i t_j}|$ is the numbers of sentences contains new phrase $t_i t_j$,

$|s_{t_i}|$ is the numbers of sentences contain keyword t_i.

A threshold θ is given to determine how many new phrases will be selected. Higher threshold will make few new phrases. In the example above, if the threshold is 0.66 then only "knowledge map" will be selected as new phase. All new phrases are tagged as nuns and repeat phrase analyze until no new phrase selected.

Formal concept analysis (FCA) is a theory that finding concept association and structure from data sets. The main characteristic of FCA is give graphical visualization between objects and attributes to users. The concept lattice not only shows the relations between concepts but also the hierarchical structure of concepts. Each sentence can see as an object and keywords as attributes and form as concept context when applying to FCA. But there was a problem when building concept lattice. The concept lattice will be a flat lattice and less hierarchical structure because of the human writing style.

The major concept of concept lattice is the most common attributes will in the top part of concept lattice and works well in classification system. But when apply to sentences, sentence contains keyword t_1 but might not contains its upper keyword t_2. For example, attribute "object" has two subordinate attributes "state" and "behavior"; but in document when mention attribute "behavior" might not refer to "object" that is why lattice become a flat lattice. Therefore, sentence must be clustered before apply to FCA. There are two major type of clustering algorithm: hierarchical and partition algorithm. In this study, we use Agglomerative Algorithms (hierarchical algorithm) as clustering method.

Each cluster applies to FCA to find out the concept lattice, the concept lattice shows the implications. For example in object oriented language documents, "software object" usually implies "method" and "attribute". The concept lattice use to expand the keyword associations implies in sentences. The keyword associations also influence the sentence weight. For example, two sentences have same weight but one sentence composed of lots weightless keyword and another composed of pairs of keyword that has relations among with. It is obvious that the sentence composed of pair keywords must have higher weight.

Complete Example of Summary Extraction

The document divided into sentences by sentence divider at first, each sentence will be marked and first and last sentence will be identified. After keyword selected, their association is analyzed by formal concept analysis and the implications is used for keyword expansion. Then the weight of each sentence is calculated according to the weight of keyword and sentence length. Finally the sentences with higher weight are collected as summary. Fig. 2 is the workflow of summary extraction.

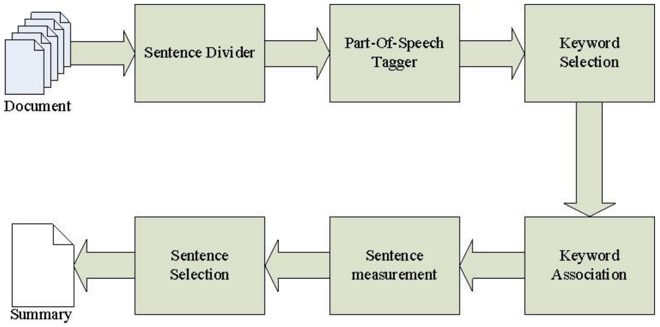

Figure 2 Summary Extraction System Work Flow

In this example, we use a document describes "What is object?" in java language. This document contains 28 sentences and after analyzed (combined keywords and remove stop world) this document contains 156 terms (words include phase). The combine rule only considers terms appear in document at last two times and the confidence must higher than 75%. There are only six new terms generated. After new term generated some words' term frequency need be modify according to the frequency of new terms such as object has been decrease by 10 because six used by "real-world object" and four used by "software object".

After new terms generated, the average lower bound of top 25% term frequency is used as selection standard. In this case, the minimum term frequency is 3 and only nouns selected as keywords. The keywords of this document are object, state, behavior, method, real-world object, bicycle, gear, software object, world, code, pedal cadence, object-oriented programming, and volume. Those keywords have been normalized by maximum term frequency. Tab. 2 is the list of term frequency and weight of each keyword.

Keyword	Term Frequency	Term Weight
Object	19	1
State	13	0.6842
Behavior	9	0.4737
Method	6	0.3158
Real-world object	6	0.3158
Bicycle	5	0.2632
Gear	5	0.2632
Software object	4	0.2105
World	4	0.2105
Code	3	0.1579
Pedal cadence	3	0.1579
Object-oriented programming	3	0.1579
volume	3	0.1579

Table 2 Term Frequency of Keyword

After the weight of each keyword is calculated, sentences are cluster into several clusters according to keywords. The concept contexts of each cluster are generated according to the result. Each concept context use formal context analysis to build their concept lattice. The implications can be found form each concept lattices and the supports also can be found. A minimum support 25% is used to filter the implications.

The document contains 13 keywords and 28 sentences; the weight of keyword expansion is assigned to $w_e = 0.5$. After clustered and FCA processes the weight of each sentence was:

$$ws = R \times W_k^T = [1.3421 \quad 1.3948 \quad 1.9737 \quad 1.6579 \quad 2.0790 \quad 2.1316 \quad 1.3948 \quad 2.1579 \quad 0 \quad 2.1316 \quad 1.3421 \quad 0.3684 \quad 0.2105 \quad 2.1842$$
$$2.4737 \quad 2.2369 \quad 2.3948 \quad 0.4737 \quad 2.8685 \quad 1.4211 \quad 1.7105 \quad 1.3421 \quad 2.1053 \quad 1.500 \quad 1.500 \quad 1.3421 \quad 0.2105 \quad 0]^T$$

The summarizations have two kinds: abstracts and extracts. In this research the summarization method of document is extracts. Extract method must assign how many content should extract from document. If assigned 25% of document should be extracted, there are seven sentences should be extracted according to sentence weight in above example. The represent of summarization is according to the position of documents. The summarization of document is show as below.

Objects are key to understanding object-oriented technology.

Real-world objects share two characteristics: They all have state and behavior.

Dogs have state (name, color, breed, hungry) and behavior (barking, fetching, wagging tail).

Bicycles also have state (current gear, current pedal cadence, current speed) and behavior (changing gear, changing pedal cadence, applying brakes)

Identifying the state and behavior for real-world objects is a great way to begin thinking in terms of object-oriented programming.

Software objects are conceptually similar to real-world objects: they too consist of state and related behavior.

Methods operate on an object's internal state and serve as the primary mechanism for object-to-object communication.

Conclusions

In this study we developed an information retrieval based automatic summarization system. This system can extract important keywords and meaningful phases from document automatically. Then system calculated the weight of each words and use formal concept analysis to find the implication between keywords. The system expands keywords according to the implications between keyword; weight sentence by sentence length; and re-weight by document structure. Finally, system extracts the summary with user defined percentages.

The summarizer not only considers about the word-level analysis but also reflect on semantics-level. The part-of-speech tagger helps us identify how each word use in sentence and through word combination we can find out phase as meaningful keyword. The formal concept analysis, we can find out the subordinate relation between keywords form sentences. Those relations indicate some semantic and can be use for further analysis. The summarizer got good result in summarize document but it still have some shortcoming to improve. In some documents, the organization is not obviously, the summarizer must analyze further. In further research, the summarizer could try to reorganized sentences into different text fragments and find out which fragments can represents document.

During the extract process, the implications between concepts were produced form formal concept analysis. Those implications indicate the subordination between keywords from text. Furthermore, collect the implications together to form as graph it becomes some kind of conceptual graph. The graph have some incorrect relations comes from writing style but it still can use to explain the content of document. With modeling technology such as Interpretive Structural Modeling (ISM), we can try to find out the hierarchy structure of keywords. By analyses sentence in text, we can adjust hierarchy structure and sent to sentence generator to generate sentences. In future, we will try to develop a system not only generate summary but also generate knowledge structures.

Acknowledgement

This study was supported by the National Science Council, Taiwan, under the Grant No. NSC98-2511-451-002.

Reference

Amini, M. R., Usunier, N. and Gallinari, P. (2005). “Automatic Text Summarization based on Word-Clusters and Ranking Algorithms.” in European Conference on Information Retrieval. Santiago de Compostela, Spain. 142-156

Cardie, C. (1997). “Empirical Methods in Information Extraction.” AI Magazine 18(4):5-79

Daumë III, H. and Marcu, D. (2006). “Bayesian Query-Focused Summarization.” Proceeding of 21st International Conference on computational Linguistics and 44th Annual Meeting of the CAL, Sydney. 305-312

Hovy, E., and Lin, C.-Y. (1999). “Automated Text Summarization in SUMMARIST.” Advances in Automatic Text Summarization, MIT Press.

Li, S.-C. (2000). “Term Association-based Hypertext Information Retrieval.” Depart of Information and Computer Engineering, Chung Yuan Christian University. Master Thesis.

Li, S.-C., Chang, J.-C. Chang, M. and Heh, J.-S. (2006) “Applying Knowledge Map to Diagnose Students’ Misconception and Provide Suitable Teaching Materials.” WSEAS Transactions on Computers 1(5): 133-140

Luhn, H. P. (1958) “The Automatic Creation of Literature Abstracts.” IBM Journal of Research and Development 2(2): 159-165

Wolff, K. E. (1994). “A First Course in Formal Concept Analysis.” Proceedings SoftStat’93, Gustav Fischer Verlag, Stuutgart. 429-438