

A Computing Model for Concept Fusing and Document Classification

Nan Zhang and Chao He

China Knowledge Grid Research Group, Key Lab of Intelligent Information Processing
Institute of Computing Technology, Chinese Academy of Sciences
P.O.Box 2704-28, 100080, Beijing, China

zhangnan@kg.ict.ac.cn, hc@kg.ict.ac.cn

Abstract

Effective document classification is a long-pursued goal in knowledge management. This paper proposes a novel hybrid approach of semantic representation and statistical measurements. Document is divided into content segments first. By Formal Concept Analysis (FCA), their semantic links with standard concept identifiers are built up whose weights are calculated statistically. In this way, effective concept fusing and document classification can be achieved. In addition, a semantic overlay for specific documents will be constructed via concept fusing. Experiments show our approach is feasible and effective.

1. Introduction

Currently, there are two basic ways for knowledge representation: 1) semantic representation and reasoning, the primary style of human cognition; 2) statistical measurements, retrieving knowledge from a large amount of documents. Formal Concept Analysis (FCA) [9], ontology [8], semantic web [2] and knowledge grid [13] are in the first category, while text mining [7], knowledge management [5] and information retrieval [3] are from the angle of statistics.

However, neither approach is effective or complete enough to deal with the classification problem, like the two sides of a coin. Although both FCA and ontology provide comprehensive conceptualization and flexible structure of domain knowledge to express semantics, neither is machine understandable. Furthermore, they can not generalize massive information into knowledge. In contrast, statistical measurements, such as data mining techniques and Bayesian analysis, utilize mathematical tools to discover frequent patterns and compose rules, and generalize knowledge schema. Concept hierarchical clustering is subdivided into agglomerative method and divisive method which fuse or refine groups respectively. They can be implemented at the machine understandable level.

Nevertheless, they are deficient in knowledge reasoning and human cognition.

Therefore, it sounds reasonable to take advantage of both semantic representation and statistical measurements. Some papers make contributions to reasoning enhancement by combining FCA and statistical analysis [11, 12]. [6, 10] adopt statistical analysis to implement large ontologies merging. In [4], concept hierarchies are agglomerated by clustering process.

The main contribution of our paper is a novel computing model for document clustering and classification. It adopts Formal Concept Analysis (FCA) for document representation. Furthermore, it extends the FCA with Bayesian estimate. By such a probabilistic structure, the model depicts the semantic overlay for documents. According to FCA, concept hierarchies of documents are organized in graphic lattice. In our model, the association relationship is bijective and probability weighted. For new documents, the approach classifies them according to such probabilistic weighted graphic lattice. Furthermore, such classification can also be formulated by undirected bipartite graph with Bayesian estimate. The model proposes two methods to classify the concept identifiers of new documents.

2. Background

Formal Concept Analysis (FCA) is an important tool for expressing the relationships between two sets of concepts [9]. Concept set is called *formal concepts*, including *formal objects* and *formal attributes*.

Definition 1. Suppose the object set O and the attribute set A , formal concept $C = \{(O_i, A_j) \mid O_i \in O, A_j \in A\}$. For formal concept C_p and C_q , $C_p \leq C_q$, if and only if $O_p \leq O_q$, or $A_q \subseteq A_p$.

Figure 1 depicts *Galois connection and graphic lattice representation of vehicles*. In (a), row is object and column is attribute. In (b), node is object or attribute, and line is conceptual relationship.

	fuel	wheel	engine	terrene	machine
auto	×	×	×	×	×
bicycle		×		×	×
carriage		×		×	
steamboat	×		×		×

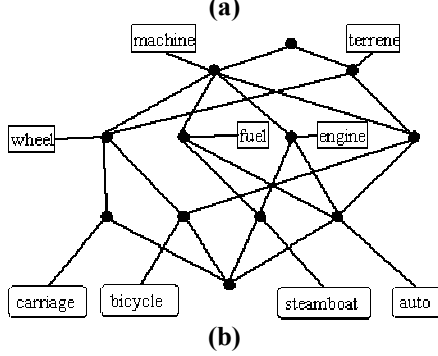


Figure 1. (a) Galois connection (b) Graphic Lattice

3. System Architecture

The aim of the computing model is to construct a probabilistic structure of concepts and documents. Herein, we adopt ACM's Computer Classification System (CCS) as the standard concept identifiers [1].

Figure 2 shows the general architecture of the computing model. It works as follows:

- 1) **Document modeling.** Based on FCA, documents are separated into content segments. The model selectively extracts semantic tags from content segments. Documents are asserted with concept identifiers according to CCS. Then, the model builds graphic lattice between concept identifiers and semantic tags. The edges in graphic lattice represent probabilistic relation.
- 2) **Concept fusing.** Concept identifiers are described in linear functions. Documents are variables in the function. The coefficients of variables are computed from probabilistic structure of concept identifiers and semantic tags. Concept fusing approach is implemented by computing the correlation coefficients among concept identifiers. By concept fusing, a semantic overlay is built upon documents.
- 3) **Document classification.** During previous two phases, semantic interconnection is integrated into graphic lattice. For an unclassified document, the model uses two algorithms (probabilistic versus graphic) to classify it into appropriate concept identifier. Finally, the model combines the results of the two algorithms and determines the optimal concept identifiers for documents.

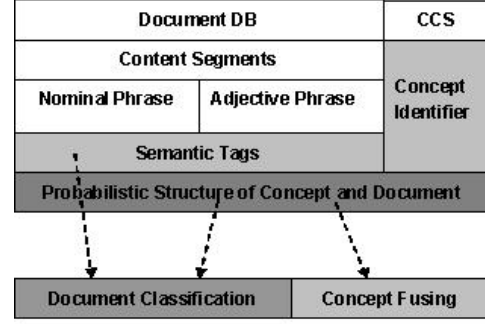


Figure 2. An overview of the system architecture

4. Main Approach

4.1 Document Modeling

Scientific papers have a certain logical structure. The main idea of a paper often resides in title, abstract and keywords. The computing model will extract these content segments to represent the knowledge embodied in documents. Furthermore, the content segments are converted into nominal or adjectival phrases, which are called semantic tags in this paper. The task of document modeling is to construct probabilistic structure between concept identifiers and semantic tags.

Definition 2. For document D and concept identifiers c_i , $c_i \in C$, $D^C = (D, C, \varphi)$, $\varphi = D \times C$.

φ defines the probabilistic relation between D and c_i , because the concept identifier assignment is subjective. But most convincing assignments can be determined according to their probability distribution. φ is calculated as follows:

$$\forall c_i \in C, \forall d_j \in D, d_j = (k_{d_j}, r_{d_j}),$$

$$\varphi(C \times D) = P((c_i \times k_{d_j}) \cup (c_i \times r_{d_j})) \quad (1)$$

By Bayesian theorem, it is rewritten as follows:

$$P(c_i \times k_{d_j}) = P(k_{d_j} | c_i) = \frac{P(c_i | k_{d_j}) \cdot P(k_{d_j})}{P(c_i)} \quad (2)$$

The probabilistic structure of keywords and concepts forms the matrix Γ_D :

$$\Gamma_D = \begin{bmatrix} P(k_{11} | c_1) & P(k_{12} | c_2) & P(k_{13} | c_3) & \dots & P(k_{1m} | c_m) \\ P(k_{21} | c_1) & P(k_{22} | c_2) & P(k_{23} | c_3) & \dots & P(k_{2m} | c_m) \\ P(k_{31} | c_1) & P(k_{32} | c_2) & P(k_{33} | c_3) & \dots & P(k_{3m} | c_m) \\ \dots & \dots & \dots & \dots & \dots \\ P(k_{n1} | c_1) & P(k_{n2} | c_2) & P(k_{n3} | c_3) & \dots & P(k_{nm} | c_m) \end{bmatrix}_{m \times n} \quad (3)$$

where each element in Γ_D is the prior probabilities of c_i divided by k_{ij} . This structure supervises concept fusing and document classification.

4.2 Concept Fusing Approach

When a document is assigned with two major concept identifiers from the probability distribution, some inner semantic connection may exist between them. By concept fusing, the correlation extent between different concept identifiers can be calculated.

Each element in Γ_D is the prior probability of c_i divided by k_{ij} , which is defined as $P(k_{ij} | c_i)$. The probabilistic contributions to $P(k_{ij} | c_i)$ can be divided into two patterns:

- 1) **Subjective definition**, for document D , $k_{ij} \in D$, if D is asserted into concept identifier c_i , k_{ij} is also asserted into the same concept identifier.
- 2) **The inherent semantics of document D** , if we consider many documents having the same k_{ij} , we can find k_{ij} are usually associated with one or more concept identifiers.

Based on the above observation, if we use $P(k_{ij} | c_i)$ to re-estimate the relation between documents and concept identifiers, the agglomerate of inherent semantics of documents can be obtained.

Suppose document d_j , $d_j \in D$ with keyword set K_d . Using the correspond element $P(k_{dj} | c_i)$ in Γ_D , the approach computes the linear expressions for document D :

$$\Delta D = \sum_i \sum_j P(k_{dj} | c_i) \cdot c_i \quad (4)$$

where c_i are independent from each other. In this way, we get the matrix of documents by concept identifiers. It can be transposed to the matrix of concept identifiers by documents, where each concept identifier is a linear expression of all the documents.

$$\Delta C = \sum_j \sum_i P(k_{dj} | c_i) \cdot d_j \quad (5)$$

The correlation level between c_i and c_j is defined as follows:

$$\text{corr}(\bar{c}_i, \bar{c}_j) = \frac{\bar{c}_i \bullet \bar{c}_j}{|\bar{c}_i| \cdot |\bar{c}_j|} \quad (6)$$

$\text{corr}(c_i, c_j)$ characterizes the degree of concept fusing. Set threshold δ , $\delta \in [0, 1]$. if $\text{corr}(c_i, c_j) \geq \delta$, c_i and c_j can be fused to one latent concept. Thus, the documents related with c_i and c_j form a semantic overlay.

4.3 Document Classification Approach

The probabilistic structure of document-concept model also supervises documents classification approach. The classification approach can be described as follows: Suppose unclassified document d_t has keywords set $d_t(k_{t1}, k_{t2}, \dots, k_{tm})$, the approach constructs a limited concept identifier set C' . C' determines the concept hierarchy upon d_t with most probability.

This paper proposes two algorithms for document classification approach: (1) posterior probabilistic estimate (PPE); (2) maximum matching of weighted bipartite graph (MWbG). After using the two algorithms, the proposed model selectively asserts the proper concept identifier for unclassified document.

4.3.1 PPE algorithm

This algorithm changes probabilistic structure of document-concept model, and computes posterior probability for each k_{tj} in unclassified document d_t . Equation 2 can be converted as follows:

$$P(c_i | k_{tj}) = \frac{P(k_{tj} | c_i) \cdot P(c_i)}{P(k_{tj})} \quad (7)$$

where $P(k_{tj} | c_i)$ is an element in Γ_D , $P(c_i)$ and $P(k_{tj})$ are the probabilities that c_i and k_{tj} appears in concept identifier set C and keyword set K respectively.

For each keyword k_{tj} in unclassified document d_t , Equation 7 estimates the probabilities that each concept identifier c_i divides keyword k_{tj} .

For $d_t(k_{t1}, k_{t2}, \dots, k_{tn})$, Equation 7 changes into:

$$P(c_i | k_{t1}, k_{t2}, \dots, k_{tn}) = \frac{P(k_{t1}, k_{t2}, \dots, k_{tn} | c_i) \cdot P(c_i)}{P(k_{t1}, k_{t2}, \dots, k_{tn})} \quad (8)$$

Suppose keywords are independent, $P(k_{t1}, k_{t2}, \dots, k_{tn})$ in Equation 8 is changed into the combination of multiple independent $P(k_{tj})$. Thus, equation 8 is reduced as follows:

$$P(c_i | k_{t1}, k_{t2}, \dots, k_{tn}) = \frac{\prod_{j=1}^n P(k_{tj} | c_i) \cdot P(c_i)}{\prod_{j=1}^n P(k_{tj})} \quad (9)$$

Equation 9 estimates all the probabilities that concept identifier set C divides unclassified document d_t . After the estimation process, the approach lists all the concept identifiers in descending order of $P(c_i | k_{t1}, k_{t2}, \dots, k_{tn})$. The approach selects the top two concept identifiers, and asserts them as the categories of d_t .

Note that PPE algorithm assumes that keywords are independent. Actually, such assumption is not always true. For example, one paper has keyword *knowledge* as well as *knowledge grid*. Apparently, *knowledge grid*

is sub-concept of *knowledge*. So, those two keywords cannot be treated independently. If we still use Equation 9 for document classification, the value of denominator will improve, while the effect that concept identifier divides document will be weakened. Our approach also proposes a graph method to find the optimization of concept identifiers to match the maximum of keywords set.

4.3.2 MWbG Algorithm

Due to the drawbacks of PPE algorithm, the approach also constructs a non-statistical method to accomplish the same task. MWbG algorithm treats concept identifiers set C and keywords set K as a bipartite graph. The edges of this graph correspond to elements in Γ_D .

Definition 3. Suppose unclassified paper d_{uc} has keywords set K_{uc} , $K'_{uc} = K_{uc} \cap K$, where K is set of row vectors in Γ_D . C'_{cr} is set of column vectors corresponding to K'_{uc} . Set G is bipartite graph, $G = (K'_{uc}, C'_{cr}, E)$ if K'_{uc} and C'_{cr} are two disjoint sets of vertices and E is a set of edges connecting vertices between K'_{uc} and C'_{cr} . Every edge in E is incident to two vertices belonging to different sets. $\forall e, e \in E, e = (k', c'), k' \in K'_{uc}, c' \in C'_{cr}$. The probabilistic elements in Γ_D are the weights of E .

Using G , the task above is deduced as finding the optimization concept identifier set C'_{cr} to match the maximum of keywords set K'_{uc} . The optimistic function can be defined as follows:

$$\forall k'_j \in K'_{uc}, \max \left\{ \frac{\sum_i \sum_j P(k'_j | c'_i)}{|c'_i|} \right\}, \quad (10)$$

where $|c'_i|$ is the number of concept identifiers in C'_{cr} .

5. Prototype and Simulation Analysis

In this section, we describe a prototype to explain our approach. Table 1 describes the relationship between papers and keywords as the training set. Table 2 shows the classification labels of the papers according to CCS. By matrix multiplication of Table 1 and 2, the relation Δ_{k-c} between keywords and concept identifiers can be calculated, where row is keyword, column is concept identifier and each element is the frequency that c_i classifies k_j .

	k_1	k_2	k_3	k_4
d_1	×		×	
d_2		×	×	×
d_3		×	×	
$P(k_j)$	$1/4$	$2/4$	$3/4$	$1/4$

Table 1. Paper-keyword mapping

	c_1	c_2	c_3
d_1	×	×	
d_2		×	
d_3			×
$P(c_i)$	$1/4$	$2/4$	$1/4$

Table 2. Paper-concept mapping

$$\Delta_{k-c} = \begin{bmatrix} 1 & 1 & \\ & 1 & 1 \\ 1 & 2 & 1 \\ & & 1 \end{bmatrix}$$

By Equation 2, the model estimates the prior probabilities Γ_D that concept identifiers divide keywords.

$$\Gamma_D = \begin{bmatrix} 0.29 & 0.14 & \\ & 0.29 & 0.58 \\ 0.43 & 0.43 & 0.43 \\ & 0.29 & \end{bmatrix}$$

Concept Fusing Recall Equation 3 and 4, the linear function of papers in training set with concept identifier variables is listed in Table 3.

		Concept Identifiers		
		\vec{c}_1	\vec{c}_2	\vec{c}_3
Δd_i	d_1	0.72	0.57	0.43
	d_2	0.43	1.01	1.01
	d_3	0.43	0.72	1.01

Table 3. Linear function of paper

Equation 6 forms semantic overlay of concepts. The correlation coefficients of concept identifiers upon training set are represented in Table 4.

	\bar{c}_1	\bar{c}_2	\bar{c}_3
\bar{c}_1			
\bar{c}_2	0.897		
\bar{c}_3	0.838	0.255	

Table 4. Correlation coefficients of concept identifiers

Set the threshold of concept fusing δ to be 0.8. The model concludes two semantic overlay c_1c_2 and c_1c_3 . Semantic overlay means that c_1 has strong correlation with c_2 and c_3 , while the semantic relation may be different between c_2 and c_3 .

Paper Classification Given an unclassified paper $d_{uc}(k_1, k_2, k_3)$, Table 5 depicts the classification result. According to PPE, d_{uc} is labeled as c_3 and c_1 , while d_{uc} is classified into $\{c_1, c_3\}$ by MWbG. Therefore, the model asserts that c_3 is the concept description for d_{uc} .

Concept identifier Set	PPE	MWbG
$\{c_1\}$	0.031	N/A
$\{c_2\}$	0.004	N/A
$\{c_3\}$	0.122	N/A
$\{c_1, c_2\}$	N/A	0.505
$\{c_1, c_3\}$	N/A	0.65
$\{c_2, c_3\}$	N/A	0.575
$\{c_1, c_2, c_3\}$	N/A	0.43

Table 5. Results of PPE and MWbG

6. Experimental Results

Our computing model is one of the basic components in knowledge management system. It facilitates research activities of computer scientists by tracking hot research topics in designated areas intelligently and classifying papers into appropriate concept identifiers automatically. We select 37 papers in the area of semantic grid as the training set. Table 6 lists the category of each paper according to CCS which is also used as the Paper ID.

ID	CCS Categories
C.2	Computer Communication Networks
H.3	Information Storage and Retrieval
K.4	Legal Aspects Of Computing
H.5	Miscellaneous
H.1	Models and Principles
D.2	Software Engineering
H.2	Database Management
K.3	Computers and Education

Table 6. The categories of CCS

The computing model extracts the keywords of documents in training set, and builds probabilistic keyword-concept structure. Figure 3 illustrates the keyword set from training set.



Figure 3. Keywords list in training set

Table 7 is the result of concept fusing. If δ is 0.8, the model estimates that concept identifier C.2 and H.3, K.4 and H.5 have strong correlations. We can conclude that the research fields of semantic grid, computer communication networks, and information storage and retrieval are much related to each other. Figure 4 describes the classification result of a paper with the keywords *semantic web*, *semantic link* and *retrieval* using PPE algorithm. The concept identifier *information storage and retrieval* best asserts the research paper with those keywords.

	C.2	H.3	K.4	H.5	H.1	D.2	H.2	K.3
C.2								
H.3	0.951							
K.4	0.515	0.592						
H.5	0.010	0.293	0.845					
H.1	0.061	0.080	0.766	0.426				
D.2	0.030	0.031	0.078	0.705	0.620			
H.2	0.1718	0.182	0.027	0.020	0.707	0.394		
K.3	0.026	0.038	0.009	0.011	0.011	0.569	0.160	

Table 7. The correlation coefficients of training set



Figure 4. Concept identifiers by PPE

7. Conclusion

This paper proposes a sophisticated computing model to exploring semantic structure of documents for intelligent knowledge management. It improves the traditional document classification based on semantic representation, and builds classifiers to automatically assert documents into appropriate categories based on document analysis. Our main contributions are:

- 1) It provides a probabilistic structure associating categories with documents resource. The model is problem oriented. It forms semantic overlays according to different document resources. Concept represented in categories can fuse into an upper concept in dynamic way.
- 2) It facilitates document classification intelligently. The classifier in the probabilistic model alleviates the subjective impact from occasionally assertion.

The proposed model is a part of the document classification mechanism in the e-science Knowledge Grid environment [13-21].

References

- [1] ACM CCS, <http://portal.acm.org/ccs.cfm>
- [2] T. Berners-Lee, J. Hendler and O. Lassila, The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities, *Scientific American*, 17, 2001, pp.4-6.
- [3] L. Cai and T. Hofmann, Hierarchical document categorization with support vector machines. *CIKM* 2004.
- [4] P. Cimiano, S. Staab. Learning concept hierarchies

from text with a guided hierarchical clustering algorithm. Workshop on Learning and Extending Lexical Ontologies at ICML-2005, Bonn, Aug, 2005.

- [5] K. Kummamuru, R. Lotlikar, S. Roy, K. Singal, R. Krishnapuram, A Hierarchical Monothetic Document Clustering Algorithm for Summarization and Browsing Search Results. *WWW* 2004.
- [6] M. Fernandez. Overview for Methodologies for Building Ontologies. In Proceedings of the IJCAI-99 Workshop on Ontologies and Problem-Solving Methods(KRR5), Stockholm, Sweden, August 1999.
- [7] B. Fortuna, D. Mladenic, and M. Grobelnik, Semi-automatic construction of topic ontology. *SIKDD* 2005.
- [8] A. G. Philpot, M. Fleischman, and E. H. Hovy. Semi-automatic Construction of a General Purpose Ontology. *International Lisp Conference*, New York. Oct. 2003.
- [9] U. Priss, Formal Concept Analysis in Information Science Cronin, Blaise (ed.), *Annual Review of Information Science and Technology*. Vol 40, 2006, p. 521-543.
- [10] S. Sekine, K. Sudo, T. Ogino, Statistical Matching of Two Ontologies. Published in the Proceedings of the SIGLEX99: Standerdizing Lexical Resources 1999; Maryland USA
- [11] M. Siff and T. Reps, Identifying modules via concept analysis, *IEEE Transactions on Software Engineering*, vol. 25, pp. 749--768, Nov-Dec 1999.
- [12] G. Snelling and F. Tip. Reengineering class hierarchies using concept analysis. In *Proc. SIGSOFT Symposium on Foundations of Software Engineering*. ACM, 1998.
- [13] H. Zhuge. *The Knowledge Grid*, World Scientific Publishing Co., Singapore, 2004.
- [14] H. Zhuge, A knowledge grid model and platform for global knowledge sharing, *Expert Systems with Applications*, 22 (2002) 313-320.
- [15] H. Zhuge, China's E-Science Knowledge Grid Environment, *IEEE Intelligent Systems*, 19(1) (2004) 13-17.
- [16] H. Zhuge, Clustering Soft-Devices in Semantic Grid, *IEEE Computing in Science and Engineering*, 4 (6) (2002) 60-62.
- [17] H. Zhuge, Semantic Grid: Scientific Issues, Infrastructure, and Methodology, *Communications of the ACM*. 48 (4) (2005)117-119.
- [18] H. Zhuge, et al, A Scalable P2P Platform for the Knowledge Grid, *IEEE Transactions on Knowledge and Data Engineering*, 17(12) (2005) 1721-1736.
- [19] H. Zhuge, Autonomous Semantic Link Networking Model for the Knowledge Grid, Technical Report of Knowledge Grid Center, KGRC-2006-01, Jan., 2006. <http://www.knowledgegrid.net/TR>.
- [20] H. Zhuge, Discovery of Knowledge Flow in Science, *Communications of the ACM*, 49 (5) (2006) 101-107.
- [21] H. Zhuge, P. Shi, Y. Xing and C. He, Transformation from OWL Description to Resource Space Model, 1st Asian Semantic Web Conference, Beijing, China, Sept. 3-7, 2006. LNCS 4185, pp.4-23.