Knowledge Organization Through Statistical Computation: A New Approach

Yang Xu and Alain Bernard

Institut de Recherche en Communications et en Cybernétique de Nantes,
Ecole Centrale de Nantes, 1, rue de la noë, bp 92101, 44321 nantes cedex 3, france
<Yang.Xu@irccyn.ec-nantes.fr> and <Alain.Bernard@irccyn.ec-nantes.fr>

Yang XU has been a Ph.D. candidate at Ecole Centrale de Nantes (France) since 2007. He received his bachelor's degree in engineering from Huazhong University of Science and Technology (China) in 2004, and master's degree in science from Peking University (China) in 2007. His research interests include knowledge management, knowledge engineering, and modeling.

Alain Bernard graduated in 1982 and obtained his Ph.D. in 1989. As an assistant-Professor, he worked 1990-96 in Ecole Centrale de Paris on product, technology and process modeling. From September 1996 to October 2001 he has been Professor in CRAN, as the head of the mechanical and production engineering team. His main research topics are related to RE, KBS, CAPP, product and process modeling, integration of economical and human aspects. His actual position is Professor and Deputy Director for Research at Ecole Centrale de Nantes, and in IRCCyN he is head of the "Virtual Engineering for industrial engineering" project.

Xu, Yang and Bernard, Alain. **Knowledge Organization Through Statistical Computation: A New Approach.** *Knowledge Organization, 36(4)*, 227-239. 30 references.

ABSTRACT: Knowledge organization (KO) is an interdisciplinary issue which includes some problems in knowledge classification such as how to classify newly emerged knowledge. With the great complexity and ambiguity of knowledge, it is becoming sometimes inefficient to classify knowledge by logical reasoning. This paper attempts to propose a statistical approach to knowledge organization in order to resolve the problems in classifying complex and mass knowledge. By integrating the classification process into a mathematical model, a knowledge classifier, based on the maximum entropy theory, is constructed and the experimental results show that the classification results acquired from the classifier are reliable. The approach proposed in this paper is quite formal and is not dependent on specific contexts, so it could easily be adapted to the use of knowledge classification in other domains within KO.

1.0 Introduction

Knowledge organization (KO) mainly concerns itself with issues of arrangement and classification of what we know so as to make it easier to communicate and understand, and to facilitate the use of knowledge. In a narrow sense, KO is considered to be a component sub-discipline within Library and Information Science (LIS), including information and management of bibliographical records, catalogues, bibliometrics such as different citation indexes, etc. In a broader sense, KO has an interdisciplinary perspective covering linguistics, mathematics, philosophy, psychology, cognitive science, sociology and management science, and its application domains include information retrieval (IR), knowledge management (KM), etc. At the deepest level, the studies about KO methods imply an epistemological discussion (Hjørland 2003). With the fast development of KO and the interchange of different disciplines, KO will continuously expand the elements of its broader field and play a more important and broader role in the future.

The task of organizing and classifying knowledge is conventionally carried out by librarians and different strategies are applied, for example documents are arranged in alphabetical order, knowledge is classified according to subject area using different numbering schemes (Svenonius 2000), etc. However, library classification systems might be inadequate for some purposes.

One main problem is how to classify newly emerged interdisciplinary knowledge. Consider a management information system as an example. Is it to be classified in the domain of computer science, management science or library science? A wise answer may be "all," or "none" and it will belong to a new category: MIS (Management Information System). But what if a physical classification system demands classification in just one existing category, e.g. a player can only represent one team in a championship?

Another problem is how librarians can learn the classification process. Some newly emerged scientific topics and works are so profound that only experts can define their proper categories, but the experts are not always there to assist the librarians. Experts, therefore, should clearly explain their deciding processes when making assignment decisions so that librarians can learn how to classify correctly. But can all the deciding processes be explained explicitly? Do all librarians have to learn all these, even though this seems an impossible task? Not only librarians, but all people who have to deal with knowledge classification problems face these two problems, including automatic classification systems that depend on rule base or keywords.

In order to solve these problems, one attempt is made to introduce the concept of probability and the process of statistical training. Probability may convert traditional classification results into a new sort of answer, e.g. computational linguistics belongs to the category of computer science with a 60% probability. This can avoid confusion when assigning interdisciplinary knowledge to one existing category. Statistical training can integrate the classification abilities of experts into a statistical model automatically without explicit explanations. This paper will reveal a new knowledge classification approach by constructing a classifier based on the maximum entropy principle. Such a classification scheme based on statistical approaches varies from former rule-based classification methods and thus brings a different point of view to knowledge classification.

2. A computational model for knowledge classification

Knowledge classification has been a challenging research topic for many years and is not simply the movement of knowledge to proper classes; rather, it is an interdisciplinary activity of placing knowledge into context. In other words, knowledge classification is a bridge between knowledge and context. Different classification schemes are deployed for different purpose,s such as the string-matching algorithm to classify the digital documents from the Compendex database (Golub et al. 2007), bibliographic classification schemes for web indexing (Dal Porto and Marchitelli 2006), and the interesting classification criterion which mainly considers the transferable and applicable degree of knowledge proposed by Novins and Armstrong (1999).

It is important to realize the fact that knowledge is always about something specific, so in KO no general knowledge can replace specific knowledge. Different fields have different KO structures and principles determined or approved by experts in those fields, and any existing knowledge organization system (KOS) reflects certain features of the domain it stands for (Ørom 2003). Consequently, as knowledge classification is always domain-oriented, this paper will take the classification of enterprise knowledge as an example so as to construct a model to explain our new approach to knowledge organization. Although the modeling process is focused on enterprise knowledge, this does not influence the adaptability of the statistical approach itself and will not obstruct it from being used extensively in other domains.

2.1 Classification criteria and element attributes

Knowledge classification is the process of assigning elements to different classes. The elements to be classified have attributes with values and the classes are determined according to a set of criteria. A set of objects can always be classified in an unlimited number of ways according to different classification criteria and attributes chosen. For example, if age is chosen as an attribute and "<18" and "≥ 18" are classification criteria, people can be classified as adult and non-adult (such classification is mainly applied in law). If age is still chosen as an attribute but the classification criteria have changed to "0-3," "4-12," "13-16," etc, people are classified as baby, child, adolescent, etc (such classification is usually adopted by sociologists or psychologists). When elements at-

tributes and classification criteria are both changed, there will be some other classification results.

As a result, the preliminary steps of knowledge classification are to determine the classification criteria and choose appropriate attributes. Smiraglia (2007) pointed out that as much work in knowledge organization is conceptual, it will be interesting to develop a concrete classification method for knowledge. Generally, the criteria chosen should be as independent on the context as possible. For example, when we classify words as commendatory or derogatory, we have already unconsciously built our language customs and cultural background into the criteria. Instead, if words are classified according to the number of letters, the gap between different customs and cultures could be negligible. However, there is a dilemma that, in practice, conceptual classifications such as "commendatory" and "derogatory" are much more useful than those with concrete criteria such as "2-letter," "3-letter," etc. As a result, a good classification system should have conceptual classes and objective criteria.

A classification model could be represented formally as follows. A set of elements $\{e_i\}$ is to be classified and n attributes are chosen to be considered, namely A_1, A_2, …, A_n, then $e_i = (a_1^i, a_2^i, ..., a_n^i)$, where a_j^i is the value of attribute A_j. A set of classes $\{C_m\}$ is described to be the homes of the elements. The classes could be described, not defined rigidly as the elements, because they should be conceptual. The classification task is to eliminate the gap between rigor and conceptualization, in other words, how to classify well-defined elements into ill-defined classes appropriately.

2.2 Problem formalization

Mai (2004) has suggested that contemporary classification research should focus on contextual information; however, a KOS generally involves a complex context of a variety of information that is relatively incomplete. In such cases, the impacts of different factors cannot be analyzed scientifically, rigidly and completely. As a result, stochastic models could be referred to, as they are fact-oriented instead of reason-oriented.

The problem of knowledge classification can be expressed as follows. Given an element $e_i = (a_1^i, a_2^i, ..., a_n^i)$, which class $C_x \in \{C_m\}$ should it be assigned to?

As discussed above, when the concept of probability is introduced, the result of knowledge classification will be determined by $p(C_x \mid a_1^i, a_2^i, ..., a_n^i)$. For example, in a KOS:

> There are four classes: C_1, C_2, C_3 and C_4;
> Three attributes are used to characterize the elements to be classified: A, B and C;
> Each attribute has several values:
> $A_1 : \alpha_1, \alpha_2$;
> $B : \beta_1, \beta_2, \beta_3, \beta_4, \beta_5$;
> $C : \gamma_1, \gamma_2, \gamma_3$
> Thus, given an element $e = (\alpha_1\beta_3\gamma_2)$, the results of $p(C_1 \mid \alpha_1\beta_3\gamma_2)$, $p(C_2 \mid \alpha_1\beta_3\gamma_2)$, $p(C_3 \mid \alpha_1\beta_3\gamma_2)$ and $p(C_4 \mid \alpha_1\beta_3\gamma_2)$ will determine which of the four classes should e be assigned to.

It is often difficult to calculate these probabilities, especially in such sophisticated contexts with insufficient information. For example, we do not know whether different constraints are independent or not, or which attribute has a higher impact during the classification process, etc. As a result, we should establish an assumption: the unbiased assumption. Explicitly, the assumption mainly contains two points about how to deal with the information that is known and unknown:

> To accept the information that is known,
> To not make any assumption or have bias about the information that is unknown.

Let us continue with the example to illustrate this unbiased assumption.

Apparently, the following constraint can be obtained:

$$\sum_{i=1}^{4} p(C_i) = 1 \qquad (1)$$

If no more information is provided, according to the unbiased assumption, we get:

$$p(C_1 \mid \alpha_1\beta_3\delta_2) = p(C_2 \mid \alpha_1\beta_3\delta_2) = p(C_3 \mid \alpha_1\beta_3\delta_2) = p(C_4 \mid \alpha_1\beta_3\delta_2) = 0.25$$

This result means that any of the four classes is equally appropriate for e.

Furthermore, if two more rules can be summarized from facts or experiences, two new constraints will be introduced into the model, e.g.:

$$p(C_2) = 4 \times p(C_1) \qquad (2)$$
$$p(C_3) + p(C_4) = 2/5 \qquad (3)$$

Again, the unbiased assumption leads to the following result:

$p(C_1 \mid \alpha_1\beta_3\delta_2) = 0.12$,
$p(C_2 \mid \alpha_1\beta_3\delta_2) = 0.48$,
$p(C_3 \mid \alpha_1\beta_3\delta_2) = p(C_4 \mid \alpha_1\beta_3\delta_2) = 0.2$

This result indicates that the most suitable class for e is class C_2.

Things will become much more complicated if more rules are found, for example, two more constraints are added as follows.

$$p(C_1 \mid \delta_2) = 0.36 \quad (4)$$
$$p(C_2 \mid \alpha_1\beta_3) = 0.6 \quad (5)$$

What will be the result for $p(C_1 \mid \alpha_1\beta_3\delta_2)$? How can we calculate the value of $p(C_1 \mid \alpha_1\beta_3\delta_2)$ or $p(C_x \mid \alpha_i\beta_j\delta_k)$ if more constraints are added due to more experience in practice and deeper comprehension of the reality?

The principle of the unbiased assumption should always be emphasized; in other words, apart from the given constraints, the probabilities should be as even as possible. The word "even" here has a meaning of "equilibration" which is usually difficult to acquire at one glance and sometimes does not even have analytical solutions. Therefore, an effective model that can integrate complexity and nonlinearity with the unbiased assumption should be referred to.

2.3 Fundamentals of the maximum entropy theory

The maximum entropy approach is a probability distribution estimation technique which is widely accepted and the kernel of the theory is avoiding bias (Jessop 2004). The main idea behind the maximum entropy is that people aim at the most uniform models that satisfy all given constraints. In more mathematical terms, the maximum entropy principle means that the probability distribution which has the maximum uncertainty should be chosen among all those that are in accordance with the available prior knowledge, i.e. a set of constraints (Kojadinovic 2007). For decades, models based on maximum entropy have been successfully applied in a variety of fields involving economic topics such as the approximation of income distribution (Wu 2003), biophysical chemistry problems (Ablonczy et al. 2003), natural language processing tasks like text classification (Nigam et al. 1999) and computational morphology (Xu and Wang 2007), manufacturing systems applications (Wang and Chuu 2004), science of materials (Böhlke 2005), etc. These applications of different fields show that the benefits of the maximum entropy framework lie in the ease of combining different information sources and knowledge into one model and the high reusability of the model. Inspiring and satisfying results are reported in most cases of its use.

Let us look at a brief introduction to the theory. As defined by Shannon (1948), the entropy is calculated as:

$$H(p) = -\sum_x p(x)\log p(x)$$

where $p(x)$ is the probability of x. If there is no information that can differentiate these xs, the best way of the least prejudiced is to regard them distributed with an equal probability, i.e. $1/n$.

The empirical distribution of x in the training set y is defined as:

$$p'(x, y) \equiv 1/N \times \text{number of times that (x, y) occurs in the sample}$$

The indicator function is introduced:

$$f(x,y) = \begin{cases} 1 \ldots \text{if y belongs to the knowledge of the class } \Omega \text{ in the context of x} \\ 0 \ldots \text{else} \end{cases}$$

The expected value of f with respect to the empirical distribution $p'(x, y)$ is:

$$p'(f) \equiv \sum_{x,y} p'(x, y) f(x, y)$$

The expected value of f with respect to the model $p(y \mid x)$ is:

$$p(f) \equiv \sum_{x,y} p'(x) p(y \mid x) f(x, y)$$

This expected value of the model is constrained to be the same as the expected value in the training sample:

$$p(f) = p'(f)$$

This equation tells us that the model can embody the statistical phenomena of the sample. If P is the space of all probability distributions, then the constraint set C, which is a subset of P, can be defined as:

$$C \equiv \{p \in P \mid p(f_i) = p'(f_i) \text{ for } i \in \{1,2,\ldots, n\}\}$$

To select a model from a set C of allowed probability distributions, we choose the model p^* with the maximum entropy $H(p)$:

$$p^* = \arg\max_{p \in C} H(p)$$

By introducing a Lagrange multiplier λ_i for each f_i, we get:

$$p_\lambda(y \mid x) = \frac{1}{Z_\lambda(x)} \exp\left(\sum_i \lambda_i f_i(x,y)\right)$$

$$\Psi(\lambda) = -\sum_x p'(x) \log Z_\lambda(x) + \sum_i \lambda_i p'(f_i)$$

where $Z_\lambda(x) = \sum_y \exp\left(\sum_i \lambda_i f_i(x,y)\right)$, which is a normalizing constant.

According to the Kuhn-Tucker Theorem (Kampas 2005), we come to an optimization problem of parameters:

$$\lambda^* = \arg\max_{\lambda} \Psi(\lambda)$$

Actually, the parameter λ_i can be regarded as the weight of the feature function, and the value of λ_i can be obtained by the training process. The computing process of parameters is quite complicated in the maximum entropy model, and in most cases, λ^* that maximizes $\Psi(\lambda)$ cannot be calculated analytically. Instead, numerical methods are applied, including the classical algorithms such as the Generalized Iterative Scaling (GIS) (Darroch and Ratcliff 1972) and the Improved Iterative Scaling (IIS) (Berger et al. 1996).

In the testing process, for a given context *x*, the decision of the knowledge classification is made according to the calculation results based on the probability distribution of the model:

$$p_\lambda(y \mid x) = \frac{1}{Z_\lambda(x)} \exp\left(\sum_i \lambda_i f_i(x,y)\right)$$

where $Z_\lambda(x) = \sum_y \exp\left(\sum_i \lambda_i f_i(x,y)\right)$

Finally, results of knowledge classification are obtained by calculating and comparing the probabilities $p_\lambda(y_m \mid x)$ of different knowledge classes.

Intuitively, the maximum entropy principle is to model what is known and assume nothing about what is unknown. In order to construct the model, a training set should be provided for model building and self-learning. In fact, the parameter computation process builds the model iteratively.

3. Application of the KO approach

This section will provide an application case in detail to illustrate the knowledge organization approach based on the maximum entropy model. Although the case mainly focuses on enterprise knowledge classification, it does not imply the application range of the approach. On the contrary, the KO approach could be applied in different domains of KO, such as library document classification, because the modeling process is similar: class description, attribute characterization, training and testing.

3.1 Class description

As for enterprise knowledge, two main aspects are usually considered in classification: cost investigated by the company, including human resources, financial support, material provision, time and opportunity, and effects gained from the knowledge. Therefore, knowledge classes can be regarded as a binary function with the knowledge cost and effect as variables, i.e. $KClass = F(\cos t, effect)$. As a result, these two variables are considered as the classification dimensions, and the enterprise knowledge can be thus classified into four classes: strategic knowledge, application knowledge, fundamental knowledge and burdened knowledge, c.f. Figure 1.

3.1.1 Strategic knowledge

Strategic knowledge lies in the region of high cost and high effect. This type of knowledge is usually rare but is playing or will play an important role in enterprise productions. However, its contribution to the enterprise is not promising and there are high risks as well. The risks can be considered as a part of cost, namely the opportunity cost. Strategic knowledge usually restricts the development of the enterprise, and high costs, including tacit costs and explicit costs, should be spent on acquiring and managing such knowledge. Development strategies of the enterprise, innovative management patterns, high technologies, brand effects, cultivation of the enterprise culture and training of the personnel with special talents and abilities belong to this class.

3.1.2 Application knowledge

Application knowledge lies in the region of low cost and high effect, and it is crucial in the current operation and management of enterprises. Such knowl-

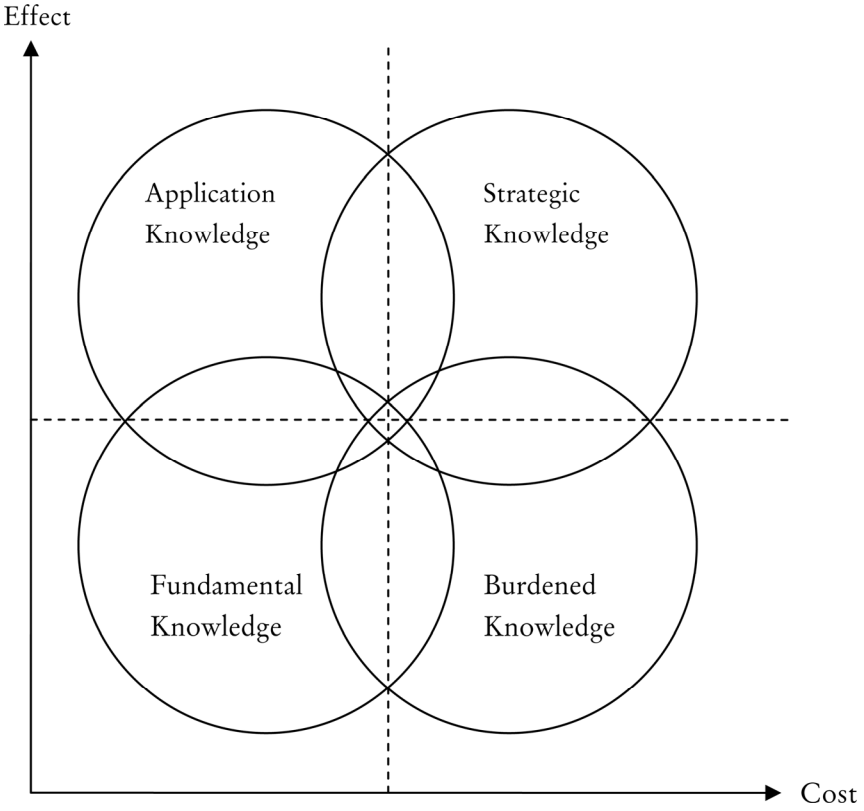

Figure 1. *Four classes of enterprise knowledge*

edge is relatively stable and mature, and can be controlled well. The application knowledge is usually in the "high value" period of the knowledge lifecycle, and it is the main source of where the profits of knowledge management come from. Existing production and management systems, application technologies, manufacturing schedules, distribution networks and corporation relationships of the enterprises belong to this class.

3.1.3 Fundamental knowledge

Fundamental knowledge lies in the region of low cost and low effect. This type of knowledge plays a supporting role in enterprise productions, including general and common knowledge in the industrial field. Fundamental knowledge is usually of high quantity and low confidentiality and can be acquired easily from the internal or external sources of the enterprise at a low cost. Laws, the macro economic status of a country or a region, statistics reports from the government, disclosed balance sheets of the cooperators or competitors, the situation of stock markets and generally applied technologies belong to this class.

3.1.4 Burdened knowledge

Burdened knowledge lies in the region of high cost and low effect. This type of knowledge is in the recessionary stage of its lifecycle and is about to quit the production system of the enterprise with its contribution diminishing. Meanwhile, relatively high cost should be spent for its general management and maintenance. Technologies that fall into disuse, out-of-date documents and abandoned businesses of the enterprise belong to this class.

However, the diversity and the dynamicity of knowledge make the classification problem inconstant. As shown in Figure 1, the four classes intercross each other which illustrates that the knowledge classification is sometimes not absolute. Instead, it will be more appropriate to say that a specific piece of knowledge has a higher probability of belonging to a certain class, and the issue about the classification probability will be discussed later. Furthermore, the class to which the knowledge belongs may be interchangeable as the knowledge life cycle is a constantly changing process. This is similar to the problem of classifying newly emerged interdisciplinary knowledge mentioned in the introduction. Such am-

biguity enhances the necessity of introducing the notion of probability.

3.2 Attribute characterization

For the practical analysis of knowledge classification, knowledge should be characterized in detail and quantitatively. Knowledge characterization is crucial for KO in both research and practical fields and knowledge in different domains emphasize different attributes. However, all kinds of knowledge could be characterized formally by means of a vector $\langle \alpha_1, \alpha_2, ..., \alpha_n \rangle$, where $\alpha_i (i = 1,2,...,n)$ are the relevant features. The case shown in this paper mainly chooses 6 attributes of enterprise knowledge for characterization from the point of view of knowledge classification. These selected features are presented in detail as follows.

3.2.1 Knowledge form

Knowledge is usually considered to have two main forms, tacit knowledge and explicit knowledge. However, in practice some knowledge has both tacit and explicit features. As there is not an absolute border between these two kinds of knowledge, it will be more appropriate to consider what is called "tacit knowledge" as a piece of knowledge that stresses the tacit aspect, and the same for "explicit knowledge". Taking a production program as an example, it may contain tacit operational strategies as well as explicit manufacturing programs. As a result, the knowledge form should not be quantified by a Boolean value but by a series of degrees, which can characterize the tacit or explicit degree of knowledge. Such degrees can be expressed by means of a ten-point scale. Generally speaking, the strategic knowledge tends to be tacit, while the fundamental knowledge tends to be explicit, and the application knowledge is between the two.

3.2.2 Knowledge granularity

Enterprise knowledge can be organized hierarchically, distributing the macro-knowledge to the micro-knowledge in a top-down way. According to the analysis of knowledge hierarchies (Levachkine 2007) and the idea of knowledge tree, the knowledge granularity is described by a knowledge tree shown in Figure 2.

As shown in Figure 2, the enterprise knowledge is organized in a tree of several levels, and each knowledge unit of a higher level consists of one or several knowledge units at its sub level. In particular, the knowledge units of the dashed line are pseudo-units which have the same attributes as their direct son nodes, and they are introduced to clarify the level number of the knowledge units. For example, K_0 represents "enterprise knowledge management," K_1 represents "operational research," K_2 represents "human resource management," K_3 represents "mathematics," and K_3 may have son nodes like "statistics," "set theory," "calculus," etc.

The knowledge granularity is determined by the level at which the knowledge unit is located, and its quantification definition is as follows.

Definition 1. If a knowledge tree has n levels in all, with the root noted as the first level, then the granularity of the knowledge unit located in the i^{th} level is equal to $(n+1-i)/n$.

The knowledge granularity defined here embodies the hierarchical situation of knowledge in the whole

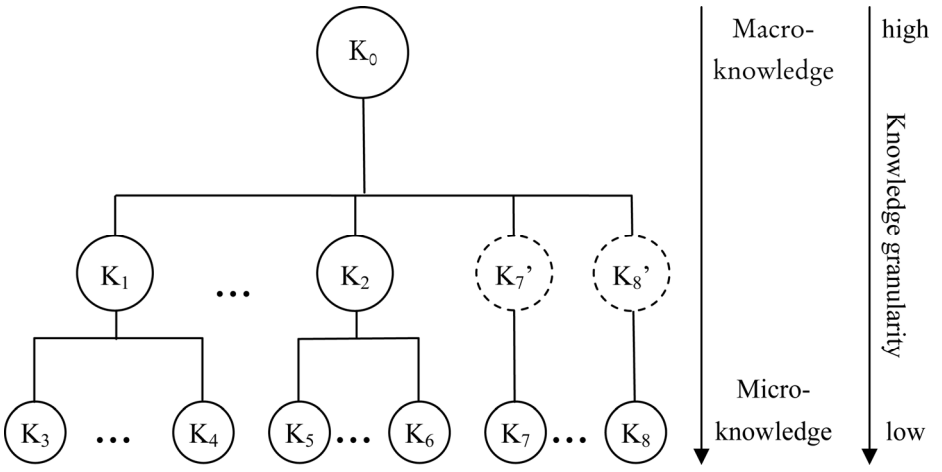

Figure 2. *Knowledge granularity described in a tree form*

knowledge organization. Generally speaking, the granularity of strategic knowledge is usually high, while the granularity of the fundamental knowledge is usually low, and the granularity of the application knowledge and burdened knowledge is in the middle.

3.2.3 Knowledge comprehensibility

Knowledge comprehensibility can be understood as "knowledge difficulty", and it characterizes the approximation of two knowledge states, the state before coding (the initial state) and the state after decoding (the object state). In most cases, there might be some differences between the expressing ability of the two states because of the process of coding and decoding. In a similar way to the knowledge form, a ten-point scale is used to measure knowledge comprehensibility. For example, an innovative scientific point of view about the cosmism is relatively difficult to comprehend and can be assigned a small number such as 2, while a simple theory about classical mechanics is easier to understand and thus can be assigned a bigger number such as 8. Generally speaking, the comprehensibility of strategic knowledge is usually low, and the application and fundamental knowledge usually have higher comprehensibility.

3.2.4 The temporal reusability of knowledge

Knowledge reuse (Majchrzak et al. 2004) is a widely discussed subject including knowledge innovation, knowledge exploring, knowledge sharing, knowledge integration, etc, and for comprehensive analysis on knowledge reuse, contexts should be considered as knowledge is context sensitive. However, the enterprise knowledge reuse issues discussed in this paper mainly refer to some cases of explicit and direct reuse which consist of two dimensions, temporal reuse and spatial reuse. This section mainly describes temporal reuse and spatial reuse will be discussed in the next section.

It is well known that the enterprise knowledge has its own lifecycle. Old knowledge retires when new knowledge generates from time to time so that the enterprises can survive in a competitive environment (Siemieniuch and Sinclair 1999). Some knowledge which is product-oriented will disappear when the product leaves the market, but some enterprise-oriented knowledge can exist throughout the whole life of the enterprise, and can last for even hundreds of years. The temporal reusability of different kinds of knowledge varies a lot, and here is its quantified definition.

Definition 2. The temporal reusability of knowledge (abbr. TRK) is equal to the time that the knowledge is used.

For example, costume design changes every year and even every season, so it is quite possible that only 5 batches of clothes with the same design are produced before the enterprise ceases such production according to the changing trend of the market. In this case, the temporal reusability of the design knowledge is 5. Here is another example that involves the knowledge of a kind of production pattern. In 1913, Mr. Ford invented the pipeline method, so the Ford Company began to use this production pattern to produce all types of cars. The knowledge of the pipeline method will not die out until other revolutionary production patterns come out, and hundreds of types of cars will be produced during its lifecycle, so the temporal reusability value of this knowledge can be up to several hundred years. Moreover, the temporal reusability value of the knowledge such as the names and the trademarks of enterprises might be $+\infty$, as it will be extremely surprising if companies like Carrefour® and Coca-Cola® will go bankrupt or abandon their brand in the predictable future. Generally speaking, the temporal reusability of strategic and fundamental knowledge is relatively high, while application and burdened knowledge has lower temporal reusability.

3.2.5 The spatial reusability of knowledge

In the enterprise system, the sharing range of different knowledge is controlled because of restrictions such as confidentiality and distribution costs, so the spatial reusability of knowledge varies greatly. Taking the design process as an example, the ideas of a designer should be limited to a specific design group rather than broadcasted to the whole company. The control on the sharing range can not only avoid the leak of commercial secrets but also prevent the employees from processing unnecessary information. Such control is quite significant in improving efficiency.

The spatial reusability of knowledge characterizes the control on the knowledge sharing range, and it has the quantified definition as follows.

Definition 3. The spatial reusability of a particular piece of knowledge has a numerical value which equals the number of these knowledge destinations.

Some examples are given to specify this definition. In a directorial group consisting of 10 managers, the spa-

tial reusability of the annual financial budget is 10. If a design drawing is to be transferred to 5 working units, the spatial reusability of it is 5. If a manufacturing parameter obtained by product analysis is distributed to 100 assembly lines, its spatial reusability is 100. Generally speaking, the spatial reusability of strategic knowledge is usually rather low because of its high confidentiality; for fundamental knowledge, its spatial reusability is relatively high as it mainly refers to general knowledge which is widely spread; application knowledge has a moderate spatial reusability.

3.2.6 Knowledge maturity

The original idea for knowledge maturity modeling is based on the levels of the CMM (Capacity Maturity Model) (Ehms and Langen 2002), and in our context, knowledge maturity describes the availability and accessibility of knowledge, and higher maturity means that the knowledge is more available and more accessible, e.g. the maturity of general knowledge is higher than that of special knowledge. As knowledge maturity is influenced by various factors and is sensitive to the context, it does not seem appropriate to give a fixed quantified definition, so a ten-point scale is used for the quantification, and different numbers will be assigned according to different enterprises, different stages and different situations. Generally speaking, the maturity of strategic knowledge tends to be low, while fundamental and burdened knowledge have higher maturity, and application knowledge is moderate.

3.3 Classification process

The enterprise knowledge classifier (EKC), based on the maximum entropy theory, is made up of a training process and a testing process. The training process mainly selects the features which have impact on knowledge classification in order to obtain a set of effective features, and to integrate these features into the model by formal descriptions, mapping algorithms and parameters computing. In the testing process, the testing knowledge is considered according to its context, and a set of effective testing features is formed by feature generation and selection. Finally, the decision results concerning the knowledge classes can be given by the EKC model.

The key tasks to implement this maximum entropy model are feature selection and the transformation of the features into model-acceptable forms. Features are selected according to the 6 attributes presented in Section 3.2, then a sextuple $\langle F,G,C,T,S,M \rangle$ is built, where F, G, C, T, S and M represent the form, granularity, comprehensibility, temporal reusability, spatial reusability, and maturity of the knowledge respectively. In addition, the features need to be generated and formalized until they become model readable, and this can be regarded as a step of information acquisition for decision making (Saunders and Miranda 1998). Among these features, some are continuous while others are discrete, and in order to avoid too many features that may cause the data sparsity problem, a mapping function is introduced, which can map all the continuous and discrete variables of the knowledge vector to a space of discrete variables. The mapping function is defined in Table 1.

In this table, the knowledge form, comprehensibility and maturity are mapped by a ten-point scale, as it is not suitable for them to be described explicitly. Although the data and the boundaries of the grades in this table may vary greatly for different types of productions, the mapping strategy and its framework can be generally applied. Thanks to this mapping function, all the selected knowledge features can be generated into the formalized features that are numbered.

Table 2 shows an example of the knowledge of a kind of selling strategy K_i.

Thus, the sextuple $(F_i, G_i, C_i, T_i, S_i, M_i)$ representing K_i is instantiated as $(4,8,3,3,4,2)$. Similarly, all knowledge pieces could be mapped to a vector, the elements of which are 6 natural numbers from 1 to 10. The training and testing process are sophisticated mathematical iterative processes shown in Section 2.3 and general software for such calculations usually provide friendly interfaces to input the model-readable vector and output the calculating results.

3.4 Experimental results

With the spreading application of the maximum entropy theory in various domains, a number of related open software is available and the software provided by Lin (http://www.cs.ualberta.ca/~lindek/maxent.tgz) is chosen as a base tool. As there are not yet any relevant standard training sets, testing sets or evaluation norms, we have chosen a certain volume of enterprise knowledge cases to form the training and testing sets, and proposed two evaluation indices based on an expert evaluation scheme (abbr. EES).

The knowledge data are provided by a medical product company and an information technology company for academic use and then distributed to the seven experts in the EES. The experts make deci-

Knowledge	Grades									
Features	1	2	3	4	5	6	7	8	9	10
Form	implicit →									explicit
Granularity	0-0.1	0.1-0.2	0.2-0.3	0.3-0.4	0.4-0.5	0.5-0.6	0.6-0.7	0.7-0.8	0.8-0.9	0.9-1.0
Comprehensibility	low →									high
Temporal reusability	<2	2-5	5-10	11-20	21-50	51-100	101-200	201-500	$501\text{-}10^3$	$>10^3$
Spatial reusability	<2	2-5	5-10	11-50	51-100	101-500	$501\text{-}10^3$	$10^3\text{-}10^4$	$10^4\text{-}10^5$	$>10^5$
Maturity	low →									high

Table 1. *The mapping function for the feature generation*

	Knowledge attributes	Description	Features	Value
K	Form	instructions written in a booklet + market situation + enterprise culture	F	4
	Granularity	at the second level of a five-level tree	G	8
	Comprehensibility	the sellers need a strong experience and good understanding of the enterprises culture to realize the idea of the strategy	C	3
	Temporal reusability	the strategy is fit for the first 10 launches of the product	T	3
	Spatial reusability	the strategy is distributed to all the 12 senior selling managers of the region	S	4
	Maturity	the strategy is not yet available to all sellers as it is still in its testing stage and not spread yet	M	2

Table 2. *An example of the data converting process*

sions on the knowledge classification independently so as to serve as "right results", and EKC will then compare its results with those "right results". This is the core idea of EES. As we know, it is impossible for humans to avoid bias, especially hidden bias (Spärck Jones 2005), so the experts are chosen from different areas with different backgrounds in order to neutralize the bias. The experts include two Ph.D.s and one engineer from USA, one Ph.D., one manager and one intern from China, and one Ph.D. from France.

To express the efficiency of the classifier, two evaluation indices, the satisfactory index (abbr. *SI*) and the comprehensive satisfactory index (abbr. *CSI*), are defined as follows.

Definition 4. The satisfactory index *SI* describes the consistency of the EKC model and the EES, and it is calculated as:

$$SI = \frac{\text{the number of situations where the EKC model makes the same decision as the experts}}{\text{the number of cases in the testing set} \times \text{the number of experts}} \times 100\%$$

Taking the subjective preference of the experts into account, the comprehensive satisfactory index *CSI* is proposed.

Definition 5. The comprehensive satisfactory index *CSI* is calculated as:

$$CSI = \frac{1}{M}\sum_{i=1}^{M}\varphi_i \times 100\%$$

$$\varphi_i = \begin{cases} 1, \text{if the result of the EKC is the same as the majority of experts in the case } i \\ 0, else \end{cases}$$

where M is the number of the test cases.

The experimental results are shown in Table 3.

As shown in the table, two kinds of tests are implemented, the 4-class test and the 3-class test. In the 4-class test, the enterprise knowledge is classified in one of the four classes: strategic, application, fundamental, and burdened knowledge; in the 3-class test, enterprise knowledge is classified in one of the three classes: strategic, application, and fundamental knowledge. The burdened knowledge class is removed in the

	4-class test	3-class test
The size of the training set	110	90
The size of the testing set	40	30
satisfactory index	65.8%	83.9%
comprehensive s satisfactory index	70.0%	90.0%

Table 3. *Experimental results*

second test because of its interference with the classification decision of the EKC model. One of the reasons for such interference may be that the burdened knowledge is not only determined by the knowledge itself but also related the production and operation states of the enterprise. In an extreme point of view, any knowledge can be regarded as burdened knowledge if needed. As burdened knowledge is of no great relevance to the enterprise production and its management method is relatively simple, it can be ignored from the consideration of knowledge classification, and it will hardly influence the effect on enterprise knowledge management. The experimental results of the 3-class test are expressed as SI=83.9% and CSI=90.0%, and they show the conformity between the EKC model and the expert group, so we can claim that the classification results of the EKC model are rather reliable and satisfying.

4. Discussions

With the fast development of newly emerged interdisciplinary domains, knowledge classification should not only rely on approaches mainly based on rules but also some methods applying statistical models which are mainly based on events. For a knowledge classification approach based on statistical computation, the core issues mainly lie in the following aspects.

1) What classification criteria to be developed? Different types of knowledge need different criteria. As the example given in the paper, a classification criterion based on cost and effect is adopted and four classes of enterprise knowledge are described according to this guideline. More specific classification schemes are to be explored in various fields; for example, in scientific document classification, the criteria could be the different intentions of the readers. Deeper discussions on classification criteria might belong to epistemology.

2) How can knowledge be characterized in order to be more beneficial to knowledge classification? Knowledge attributes and their characterization serve as the basis of all kinds of research on knowledge, and appropriate characterization strategies can facilitate relevant studies. Aiming at the issues of knowledge classification, this paper introduces the knowledge vector for characterization, which shows its efficiency and adaptability.

3) Which attributes of knowledge have higher impacts on knowledge classification and how do they affect each other? When trying to answer this question, we found that it is almost an impossible task. In the case of enterprise knowledge, it seems that logical rules are not obvious, and that is the reason why we are turning to statistical methods to apply the maximum entropy model. One of the most important advantages of statistical approaches is the integration of the sophisticated rules without showing them explicitly, just as we can say that the probability of get a 6 from a falling dice is 1/6 without telling why. It would be very interesting if we could make advances in deducing some logical rules among the impact factors and classification results. We could then construct an improved hybrid model integrating both rules and statistics.

4) How to evaluate the performance of the model? One of the most interesting features of knowledge is its subjective aspect (Raju et al. 1995), so the evaluation of the performance related to a knowledge system is sometimes subjective, at least not as objective as the evaluation of the temperature or the length. Our evaluation criterion refers to an expert evaluation scheme consisting of humans, and although the decisions made by experts may be subjective to a certain extent, it nicely reflects the subjective aspect of the decision itself. To develop a relatively more objective judging standard, more rational evaluating norms that fit the practical use should be worked out in future research.

More comprehensive studies on knowledge classification may rely on a systematic comparison of the maximum entropy approach to other state-of-the-art methods used for classification in complex systems, such as the support vector machines (Rossi and Villa 2006) and classification trees (Noh et al. 2004).

5. Concluding remarks

The classification issue in KO is the main theme of this paper. To work on this subject, three questions are raised: why classify knowledge, how to classify knowledge through a statistical approach, and whether our approach works. The aim of knowledge classification is to assign different kinds of knowledge to different classes so as to obtain a link between concrete knowledge and conceptual description. To approach this aim, we formalize the classification problem with elements (concrete knowledge) and classes (conceptual description) in a way that can also facilitate further application. Such formalization is instantiated by a specific case of enterprise knowledge classification where the characterization pattern with vectors is introduced. Based on this fundamental preparation, the paper continues the survey on knowledge classification by setting out an assumption of avoiding bias which leads to the maximum entropy theory. Then, the fundamentals of the theory are introduced and the modeling process of the classifier is analyzed in detail. In the application case, explicit data processing methods are presented, including generation and conversion of knowledge features, data collection, expert selection and a proposition of a scheme to examine performance evaluation. Finally, the experimental results show that the classifier is an effective and reliable decision supporting tool and thus illustrate that the approach for knowledge organization through statistical computation may have a promising future.

References

Ablonczy, Zsolt, Lukács, András and Papp, Elemér. 2003. Application of the maximum entropy method to absorption kinetic rate processes. *Biophysical chemistry* 104: 249-58.

Berger, Adam, Della Pietra, Stephen and Della Pietra, Vincent. 1996. A maximum entropy approach to natural language processing. *Computational linguistics* 22: 39-72.

Bernard, Alain and Tichkiewitch, Serge. 2008. *Methods and tools for effective knowledge life-cycle-management*. Springer, Berlin.

Böhlke, Thomas. 2005. Application of the maximum entropy method in texture analysis. *Computational materials science* 32(3-4): 276-83.

Dal Porto, Susanan and Marchitelli Andrew. 2006. The functionality and flexibility of traditional classification schemes applied to a content management system (CMS): Facets, DDC, JITA. *Knowledge organization* 33: 35-44.

Darroch, John N. and Ratcliff, Douglas. 1972. Generalized iterative scaling for log-linear models. *Annals of mathematical statistics* 43: 1470–80.

Ehms, Karsten and Langen, Manfred. 2002. Holistic Development of Knowledge Management with KMMM®. (available at www.kmmm.org)

Golub, Koralika, Hamon, Thierry and Ardo, Anders. 2007. Automated classification of textual documents based on a controlled vocabulary in engineering. *Knowledge organization* 34: 247-63.

Hjørland, Birger. 2003. Fundamentals of Knowledge organization. *Knowledge organization* 30: 87-111.

Jessop, Alan. 2004. Minimally biased weight determination in personnel selection. *European journal of operational research* 153: 433-44.

Kampas, Frank J. 2005. Tricks of the Trade: Using Reduce to Solve the Kuhn-Tucker Equations. *The mathematica journal* 9: 686-89.

Kojadinovic, Ivan. 2007. Minimum variance capacity identification. *European journal of operational research* 177: 498-514.

Levachkine, Serguei and Guzmán-Arenas, Adolfo. 2007. Hierarchy as a new data type for qualitative variables. *Expert systems with applications* 32: 899-910.

Mai Jens-Erik. 2004. Classification in context: Relativity, reality, and representation. *Knowledge organization* 31: 39-48.

Majchrzak, Ann, Cooper, Lynne P and Neece, Olivia Ernst. 2004. Knowledge reuse for innovation. *Management science* 50(2): 174-88.

Nigam, Kamal, Lafferty, John and McCallum, Andrew. 1999. Using maximum entropy for text classification. *Proceedings of the IJCAI-99 Workshop on Machine Learning for Information Filtering*. Stockholm, Sweden, pp. 61-67.

Noh, Hyun Gon, Song, Moon Sup and Park, Sung Hyun. 2004. An unbiased method for constructing multilabel classification trees. *Computational statistics & data analysis* 47: 149-164.

Novins, Peter and Armstrong, Richard. 1999. Choosing your spots for knowledge management. (available at: www.providersedge.com/docs/km_articles/Choosing_Your_Spots_for_KM.pdf)

Ørom, Anders. 2003. Knowledge Organization in the domain of Art Studies - History, Transition and Conceptual Changes. *Knowledge organization* 30: 128-43.

Raju, P.S., Lonial, Subhash C. and Mangold, W. Glynn. 1995. Differential effects of subjective knowledge,

objective knowledge, and usage experience on decision making: an exploratory investigation. *Journal of consumer psychology* 4: 153-80.

Rossi, Fabrice and Villa, Nathalie. 2006. Support vector machine for functional data classification. *Neurocomputing* 69(7-9): 730-42.

Saunders, Carol and Miranda, Shaila. 1998. Information acquisition in group decision making. *Information & management* 34: 55-74.

Shannon, Claude E. 1948. A Mathematical Theory of Communication. *The Bell System technical journal* 27: 397-423 & 623-56.

Siemieniuch, Carys E. and Sinclair, Murray A. 1999. Organizational aspects of knowledge lifecycle management in manufacturing. *International journal of human-computer studies* 51: 517-47.

Smiraglia, Richard P. 2007. Performance works: Continuing to comprehend instantiation. *Proceedings of North American Symposium on Knowledge Organization* (1): 75-86, Toronto, Ontario.

Spärck Jones, Karen. 2005. Revisiting classification for retrieval. *Journal of documentation* 61: 598-601.

Svenonius, Elaine. 2000. *The intellectual foundation of information organization*. The MIT Press, Cambridge, MA.

Wang, Reay-Chen and Chuu, Shian-Jong. 2004. Group decision-making using a fuzzy linguistic approach for evaluating the flexibility in a manufacturing system. *European journal of operational research* 154: 563-72.

Wu, Ximing. 2003. Calculation of maximum entropy densities with application to income distribution. *Journal of econometrics* 115: 347-54.

Xu, Yang and Wang, Hou-feng. 2007. A Hybrid Model for Computational Morphology Application. *IEEE Proceedings of the 8th International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing* (2). Qingdao, China, pp. 232-37. http://www.cs.ualberta.ca/~lindek/maxent.tgz