ADDING NEW CONCEPTS ON THE DOMAIN ONTOLOGY BASED ON SEMANTIC SIMILARITY

Rung-Ching Chen, Ya-Ching Lee, Ren-Hao Pan
Department of Information Management, Chaoyang University of Technology
168, Jifong East Road, Wufong Township, Taichung County 41349, Taiwan ROC
{crching, s9314628, s9314605}@cyut.edu.tw

ABSTRACT

In recent years, ontology application has become one of the most popular research topics on the Internet. Many researchers use domain ontology to facilitate knowledge sharing and reusing. However, the specification of the ontology is till incompletely now, due to have not a standard expression and creating method for the ontology. The information is changed fast on the Internet, so the knowledge on the ontology often can not really meet the user's requirement. Therefore, how to create a renewable ontology is necessary. In this paper, we will propose an ontology renewable method which can insert new keywords into a corresponding constructed ontology. The novel method uses LSA, Latent Semantic Analysis, to strengthen the semantic characteristic of keywords and transforms the LSA matrix to a high dimensional space based on collected web pages. The similar values between keywords and concepts are gotten by comparing the high dimension keyword to the corresponding constructed ontology concepts. If a keyword has highest similarity value between it and one of concepts on the domain and similarity value less than a threshold, the keyword will be inserted into the offspring of the concept on the domain ontology. The primary experiment results indicated the method is useful.

Keyword：Semantic Web, Domain Ontology, LSA Matrix, CKIP Segmentation

1. INTRODUCTION

Many users use the Internet technology and its application causing the information explosive growth. Consequently, users must spend a large amount of time to browse and search information. However, information is difficult to integration. In view of this, Berners-Lee et al., (1999, 2001) proposes a new network framework "Semantic web" that making the information on the network all has ability of "describe by myself". The software agent can filter and extract useful information automatically to help users to deal with and to finish every job. Hypertext links are used to organize distributed files, web contents and interface display that are designed for the humanity to browse. Due to the computer does not understand the semantic among web contents, so it is unable to deal with and filter the information for users. "Semantic web" emphasizes the structure of document on the Internet and sets up an information sharing and knowledge reusable cyberspace. That is, turn the data into the semantic network. So, the computer can understand the user's real meaning and help the user to obtain the information. The information is no longer "Machine-readable" that is turned into "Machine-understandable". Tim Berners-Lee believed the "Semantic web" can help the computer fast to find the need of user in the huge information of network. The "Semantic web" can expand human knowledge limitlessly. The computer can understand all things that linked on the network. Therefore all the things on the network become meaningful and increased the worth of the network. The Miller(2004) proposes that the final goal of the "Semantic web" is:

(1) Web of data - Offer a common data representation framework to promote integrating resources of all kinds to draw new conclusions.
(2) Increase the utility of information through linking it to its definitions and content.
(3) More efficient information access and analysis.

An important factor whether "Semantic web" can succeed is to set up attributes to the establishment of ontology. The word "ontology" was widespread quote in the Artificial Intelligence domain in recent years. A lot of definitions about ontology are being proposed constantly. Most often quoted the definition is Gruber proposed in 1993. "An ontology is a formal, explicit specification of a shared conceptualization." In definition the "conceptualization" is the abstract model of the phenomenon in existence, the word "shared" points out the ontology is shared and belonged to the collective not individual. The meaning of formal is machine can to read and understood the ontology. The meaning of explicit is used to express the concepts and limits of ontology(Gruber, 1993).

Typical ontology includes taxonomy and a set reasoning rule. Taxonomy defines the types of objects and relations between them. The clear definition of logic rules will let ontology has stronger functions. The computer will understand the meaning of web pages through linking concepts to concepts under the ontology. The computer can process humanity need information more effectively (Abasolo & Gómez, 2000). The ontology is regarded as one kind of metadata. The common ontology describing languages include: RDF, RDFS, DAML+OIL, OWL, etc. The Resource Description Framework (RDF) is a foundation of describing language. According to the definition of RDF primer of W3C, RDF is for representing information about resources on the World Wide Web. In particular to represent the metadata about Web resources, such as the title, author, and modification date of a Web page, copyright and licensing information about a Web document, or the availability schedule for some shared resource(Manola & Miller, 2004). The RDF utilizes a set of terminology to express each part of the web page statement, called *triples*. In the triples notation, each statement is written as a simple triple of subject, predicate, and object. Used for discerning that the parts of the things are called the *subject*, that can distinguish the part of different property (for instance, the author, establishing date, languages, etc.) of subject is called *predicate* and the *object* is assign to value of property. Use the example in RDF primer of W3C to illustrate, there has a statement of web page: “http://www.example.org/index.html **has a** creator **whose value is** John Smith.” The triples notation is expressed as follows(Manola & Miller, 2004):

(1) The subject is the URL http://www.example.org/index.html
(2) The predicate is the word “creator”
(3) The object is the phrase “John Smith”

Ontology not only provides the knowledge sharing and reusable, but also communicates to different agents. In addition, ontology can improve the precision of search engine too. Search engines only look for the web pages with accurate concept(Abasolo & Gómez, 2000). Users can obtain necessary information effectively based on constructed ontology. Ontology plays an important role in the semantic web. So, there have proposed a lot of construct method of ontology. Generally speaking, the ontology construction can be divided into the manual and semi-automatic. The manual construction usually carries on by the domain expert. Though this method is comparatively accurate, but it consumes too much time and manpower. The full manual construction ontology method will be relatively partial to the constructor's subjective consciousness. The major method of semi-automatically construction ontology has the following three types:

(1) Based on dictionary

The dictionary has defined the synonym, stem and root to each vocabulary. So, that can utilize the dictionary to define the relation between the vocabulary and vocabulary and decide the layer relation of concepts. But this method will be limited to the size and quality of the dictionary(Khan & Luo, 2002; Tan et al., 2000).

(2) Base on cluster of context

This method will get the similar vocabularies to group together by a clustering algorithm and select the vocabularies of higher frequency to represent this group. That can obtain the vocabulary layer after the repetition training(Williams & Tsatsoulis, 2000). Chen et al.,(2005) proposes utilizing the ART network of neural network as a cluster tool. Then this method uses conditional probability to find the related between vocabularies and to combine “Is_a” notion to judge the proper position of the concept.

(3) Base on association rule

This method utilizes association rules to select the relation between the concepts. When two concepts both have high support and confidence values, get back to the original context to look for the verb between two concepts to represent their relations(Wei et al., 2000).

However, while the ontology constructed, it is hard to add new keyword into the ontology automatically but the changing fast of information at any time, so update the ontology is needed. On the other side, with the development of information technology and new knowledge is proposed constantly. Original ontology will not suit to users’ requirement. Therefore, the problem about ontology upgrade is important. This paper will propose a novel method to do the ontology upgrade that enable ontology to accord with the description of the field and help the goal of the semantic web to reach further. First, the system uses CKIP system, Chinese Knowledge Information Processing, to do on-line Chinese words segmented service. After stop words are filtered, we set up a keyword-document matrix. Next, the LSA method, Latent Semantic Analysis, is used to strengthen the semantic characteristic of keywords to get a LSA matrix. Furthermore, the LSA matrix is transformed to a high dimensional space. Based on the high dimensional space, the similar value is gotten by comparing the high dimension keyword to the corresponding constructed ontology concepts. Finally, the keyword with highest similarity value is inserted into the ontology to modify the number of concepts on the domain ontology. The new ontology has more completely information for users' requirement than old ones.

The remainders of the paper are organized as follows. Section 2 introduces the related technologies in this research. In Section 3, we describe an algorithm to insert a keyword into an existed ontology. Section 4 specifies primary experiment results. The paper makes conclusions and future works in Section 5.

2. RELATED TECHNOLOGIES

Two major technologies, TF-IDF and LSA, are used to implement the system. TF-IDF is used to calculate the importance of the keyword in the related documents and LSA is used to find the latent information from the documents. In this section, we will introduce the two technologies respectively.

2.1 TF-IDF

Two major factors are worthily considered to determine the importance of a vocabulary in the document. The first one is the relative frequencies of occurrence with vocabulary in the document, called the Term Frequency(TF); the second is the frequencies of occurrence with document in all of document set is called the Document Frequency(DF). Term frequency measures the relative importance of a keyword in a document. If the value of the term frequency in the document (document in here pass by already remove that stop word) is very high, this term is very important that can represent this document. Document frequency measures how many documents have this term. The more the vocabulary appears in the document but in other documents few, the more this vocabulary is suitable for and used for distinguishing with other documents.

Synthesize the above two factors, we can calculate the importance of a vocabulary in a certain document by the product of term frequency and inverse of document frequency. This approach is called TFIDF in information retrieval; its formula lists as follows:

$$W_{i,j} = TF_{i,j} \times IDF_i = \frac{tf_{i,j}}{MAX(tf_{i,j})} \times \log(\frac{N}{n_i})$$

$W_{i,j}$: Denote a certain term i importance in document j.

$TF_{i,j}$: Denote relative frequency of term i appearing in document j . That is, the absolute frequency($tf_{i,j}$) of term i in the document j divided by the maximum term frequency(max($tf_{i,j}$)) of all document j.

IDF_i: Denote the inverse frequency of term i in all document set appearing.

N: Amount of the total document.

n_i: Denote the frequency of term i appearing in total documents.

2.2 Latent Semantic Analysis, LSA

Traditional information retrieval methods often use the documents and keywords match to find the keyword information. LSA is based on SVD (Singular Value Decomposition) and dimension reduction in the linear algebra with the statistics method. SVD is a decompose technology of the matrix in mathematics, that can extract knowledge with implicit of the document and transfer it into the semantic space. Dimension reduction can remove noise in the semantic space and enable the LSA accurately to find out the knowledge with implicit of the content from the document set.

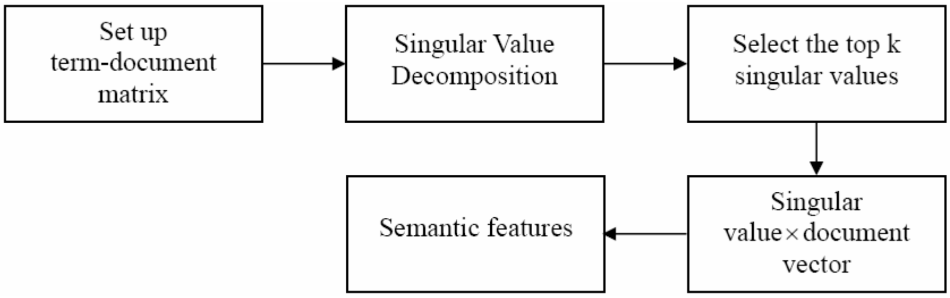

Figure 1. The workflow of LSA.

Generally the value analysis is often with matrix in mathematics. Only has a unique solution, when the matrix is a nonsingular matrix. But if the matrix is a singular matrix, the value of determinant is 0. If determinant value is 0 when being the singular matrix, it is unable to solve with general matrix. In order to solve the linear least squares and singular matrix problems, that SVD use the eigenvalue and eigenvector to reduce the dimensions of the original data and to filter irrelevant information.

The LSA flow is shown in Figure 1. Assume a terms-documents matrix X with t × d matrix(such as Figure 2), where *t* is the number of keywords and *d* is the number of documents. Each element *x* in the matrix is the number of occurrences of keyword in document. The matrix X decompose by SVD can obtain the three matrices: $X=WSP^T, S = diag(\sigma_1....\sigma_n)$, where the elements of S are all singular values of X. Let n=min {t, d}, and the singular value is represented by $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_n \geq 0$. W and P^T are t × t , d × d matrices, respectively.

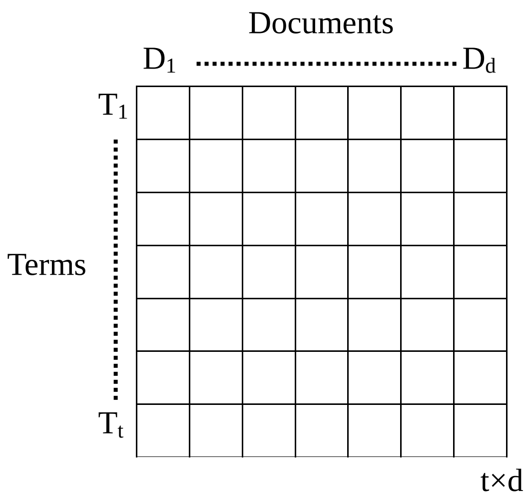

Figure 2. Terms-documents matrix.

Figure 3 shows the matrix X after SVD and dimension reduction, original matrix X simplifies to W_k, S_k and P_k^T. The dimensions of W_k, S_k and P_k^T are reduced to t × k, k × k and k × d. The common element k is less than the original vector space. W_k is a term vector. S_k retains k large singular value in terms-documents matrices. P_k^T is a document vector. Afterwards rebuild matrix to get the LSA matrix, obtain the knowledge with implicit of the document content.

$$\underbrace{\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & \ddots & & \\ \vdots & & \ddots & \\ x_{t1} & \cdots & \cdots & x_{td} \end{bmatrix}}_{X} \cong \underbrace{\begin{bmatrix} w_{11} & \cdots & w_{1k} \\ \vdots & \cdots & \vdots \\ w_{t1} & \cdots & w_{tk} \end{bmatrix}}_{W_k} \times \underbrace{\begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & \cdots & \sigma_k \end{bmatrix}}_{S_k} \times \underbrace{\begin{bmatrix} p_{11} & \cdots & p_{1d} \\ \vdots & \cdots & \vdots \\ p_{k1} & \cdots & p_{kd} \end{bmatrix}}_{P_k^T}$$

Figure 3. Matrix X was decomposed by the SVD and dimension reduction.

This research will apply the above two technologies to find out the suitable keywords to insert into the domain ontology.

3. METHOD

The workflow of new concept sets into the ontology framework shown on Figure 4. That may divide into three phases: web pages preprocessing, keywords similar comparing as well as increase the concept. We will focus on the keywords similar compare. Hope through the method of keyword similar to obtain the keyword insert to ontology.

3.1 Web pages preprocessing

In this phase, we utilize CKIP system, on-line remote Chinese words segmented service, as words segmented tool. The webpage related to ontology is collected at first. Next, the CKIP will do words segmented. The CKIP will add the vocabulary after words segmented the Part-of-speech (POS) tagger. We utilize this Part-of-speech (POS) tagged to draw out the noun. Then, remove stop words. The detail steps of web pages preprocessing include Chinese words segmented and remove stop words. Chinese sentence is continuous words. So computer is difficult to judge which words can make up a keyword. The common Chinese words segmentation has two kinds of methods roughly: dictionary word segment, statistics word segment. Dictionary word segment is the fastest and most convenient method. The Chinese Knowledge Information Processing Group offers the CKIP which is a on-line serve of words segmentation tool based on dictionary word segment. In English vocabulary like “is”, “to”, “the” and Chinese words like “你”, “了” and so on, that all call the stop words. These stop words are not meaningful in web pages analyze but has high frequency in documents so we remove the stop words from the documents to do further analysis.

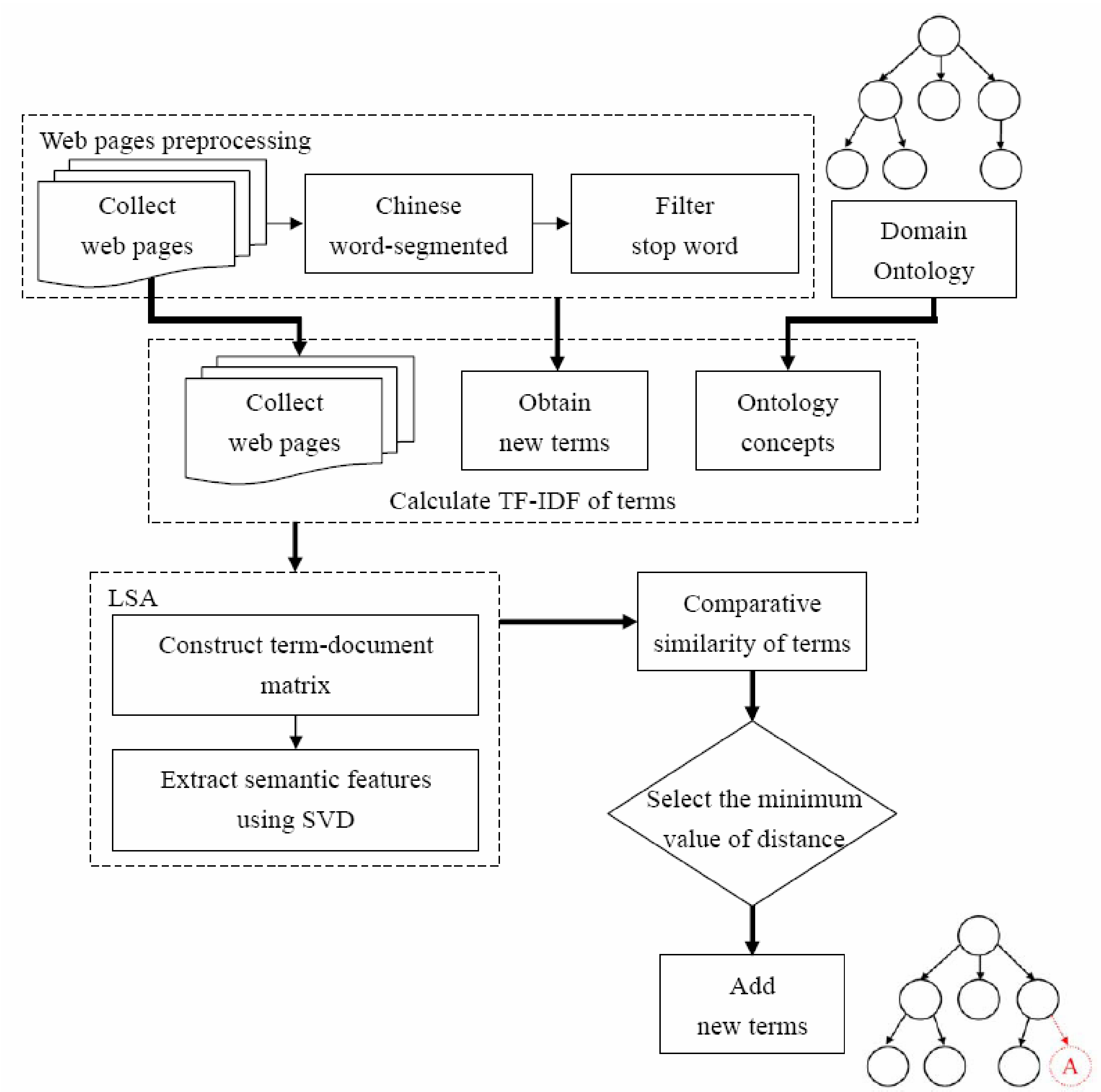

Figure 4. Insert the new concept in the ontology framework.

3.2 Finding similar values between keywords and concepts

After filtering the stop words, the system takes out the keywords to represent the related field. TF-IDF is used to calculate the importance of each keyword on the documents to receive the weight highest and fully show the keywords of this specific field. Though undergoing the stage of remove stop word, but that still has too much unallied keywords. After calculating the weight of each keyword, we will leave out the keywords with lowest weight and only keep the keywords with highest weight. The operation of TF-IDF includes extracting keywords with documents and concepts with documents. So the documents are an interface to find the relationship between keywords and concepts. Figure 5 shows the relationship of keywords and concepts. Then the highest weight keywords and document to build the keywords-documents matrix. The SVD operation decomposes the keywords- documents matrix into the three matrices and via the step of dimension reduction and rebuild matrix get a LSA matrix. LSA operation obtains the similar values between new keyword and ontology concepts. Finally, the LSA matrix will be transformed to the high dimensional space and Euclidean distance is used to calculate the distance between keyword and concepts.

$$dist(x_i, y_j) = \sqrt{\sum (x_i - y_j)^2}$$

where x_i represents a keyword i , y_j represents concept j of the ontology. The distance between the keywords more short they are more similarly.

3.3 Insert the new concept

It is similar between new concept and ontology concept in order. When finds the new concept most similar with the ontology concept, then this ontology concept will produces a branch node. Figure 5 indicates (a) If the semantic similarity value between the new concept "A " and the root is lowest and less than a threshold, the system adds a new branch node from the concept root. (b) If the semantic similarity value between the new concept "A " and the ontology concept "B" is lowest and less than a threshold, the system adds a new branch node from the concept "B". The new keyword can be added into the domain ontology automatically by above-mentioned methods. Finally, the ontology is upgraded.

(a) **(b)**

Figure 5. Adds a new keyword into ontology: (a) add a new keyword into ontology framework from root. (b) add a new keyword into ontology framework from concept B.

4. EXPERIMENT

The experiment was implemented on a Windows XP operating system using JAVA (JDK 1.4.2) with a Pentium-4 2.4G MHz CPU and 512 MB RAM. This paper uses a domain ontology covering the basketball domain using an OWL. The domain ontology includes 25 concepts and the relation of concepts is "has" or "sub-class" as defined by the taxonomy. The depth of the domain ontology is four. Its structure shows in Figure 6. The system collected news from WWW and CKIP is used to on-line Chinese words segmented service. We will detail describe the experiment steps and discusses the experimental result as follows.

4.1 Web pages preprocessing

In collection web pages stage, we collect the news web pages related sport of basketball domain. The news web pages extracted from the Yam sports (http://sports.yam.com/) under the basketball catalogue. We collect altogether 580 news web pages are date issued between 06/21/2005 and 10/11/2005. We utilize CKIP on-line Chinese words segmented service(CKIP) as words segmented tool. CKIP provides an API for users and the data exchange using XML format. Users can write the program via TCP Socket line to convey the proof information and text to the server. After server processing uses the original line to pass on the result. This on-line service is based on dictionary and exports simple POS tagger. Figure 7(a) and Figure 7(b) is an example of CKIP segmented service.

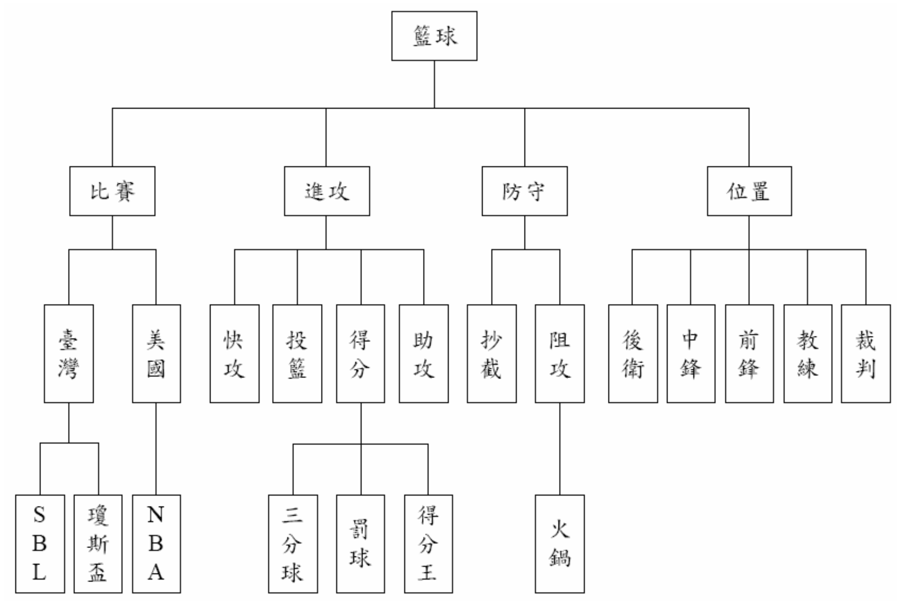

Figure 6. One of basketball ontology.

上個球季在塞爾蒂克及太陽陣營擔任替補前鋒工作的Walter McCarty，下季將改披快艇戰袍亮相。

快艇球團表示，McCarty已通過體能測試，他們準備在美國時間本週五正式與他簽約。然而，截至目前為止，快艇方面都未透露切確的合約內容。

身高6呎10吋、體重230磅的McCarty已在NBA待了九個年頭，先後效力於尼克、塞爾蒂克及太陽陣營，生涯共打過557場例行賽事，但平均上場時間只有18分鐘，每場比賽也僅有5.3分、2.7籃板及1.2助攻表現。

Figure 7(a). The source document.

。(PERIODCATEGORY) 上(Nes) 個(Nf) 球季(Nd) 在(P) 塞(VC) 爾(Nb) 蒂(Na) 克(VC) 及(Caa) 太陽(Na) 陣營(Na) 擔任(VG) 替補(VC) 前鋒(Na) 工作(Na) 的(DE) Walter(FW) McCarty(FW) ，(COMMACATEGORY)
，(COMMACATEGORY) 下(Nes) 季(Nd) 將(D) 改(VC) 披(VC) 快艇(Na) 戰袍(Na) 亮相(VA) 。(PERIODCATEGORY)
。(PERIODCATEGORY) 快艇(Na) 球團(Na) 表示(VE) ，(COMMACATEGORY)
，(COMMACATEGORY) McCarty(FW) 已(D) 通過(VC) 體能(Na) 測試(VC) ，(COMMACATEGORY)
，(COMMACATEGORY) 他們(Nh) 準備(VF) 在(P) 美國(Nc) 時間(Na) 本(Nes) 週五(Nd) 正式(VH) 與(P) 他(Nh) 簽約(VA) 。(PERIODCATEGORY)
。(PERIODCATEGORY) 然而(Cbb) ，(COMMACATEGORY)
，(COMMACATEGORY) 截至(P) 目前(Nd) 為止(Ng) ，(COMMACATEGORY)
，(COMMACATEGORY) 快艇(Na) 方面(Na) 都(D) 未(D) 透露(VE) 切(VC) 確(D) 的(DE) 合約(Na) 內容(Na) 。(PERIODCATEGORY)
。(PERIODCATEGORY) 身高(Na) 6(FW) 呎(Nf) 10(FW) 吋(Nf) 、(PAUSECATEGORY) 體重(Na) 230(FW) 磅(Nf) 的(DE) McCarty(FW) 已(D) 在(P) NBA(FW) 待(VC) 了(Di) 九(Neu) 個(Nf) 年頭(Nd) ，(COMMACATEGORY)
，(COMMACATEGORY) 先後(Nd) 效力(Na) 於(P) 尼克(Nb) 、(PAUSECATEGORY) 塞(VC) 爾(Nb) 蒂(Na) 克(VC) 及(Caa) 太陽(Na) 陣營(Na) ，(COMMACATEGORY)
，(COMMACATEGORY) 生涯(Na) 共(Da) 打(VC) 過(Di) 557(FW) 場(Nf) 例行賽(Na) 事(Na) ，(COMMACATEGORY)
，(COMMACATEGORY) 但(Cbb) 平均(VH) 上場(VA) 時間(Na) 只有(Cbb) 18(FW) 分鐘(Nf) ，(COMMACATEGORY)
，(COMMACATEGORY) 每(Nes) 場(Nf) 比賽(Na) 也(D) 僅有(VJ) 5(FW) .(PERIODCATEGORY)
.(PERIODCATEGORY) 3(FW) 分(VC) 、(PAUSECATEGORY) 2(FW) .(PERIODCATEGORY)
.(PERIODCATEGORY) 7(FW) 籃板(Na) 及(Caa) 1(FW) .(PERIODCATEGORY)
.(PERIODCATEGORY) 2(FW) 助攻(VA) 表現(VC) 。(PERIODCATEGORY)

Figure 7(b). After CKIP segment the text.
Figure 7. The operation of CKIP system.

4.2 Comparative similar between keywords

We can utilize CKIP on-line Chinese words segmented service offered POS tagger to take out the noun from the document. There are 4242 words. After removing stop words, it still has 2738 keywords. These keywords may be unallied or unimportant. Therefore, the system utilizes TF-IDF to calculate the weight of each keyword then leave out the keywords with weight less than a threshold. So the remainders 48 keywords are used to represent basketball domain. In order to obtain the keywords that semantic similar the ontology concept, these keywords with higher weight, ontology concept and documents build the terms-documents matrix shown in Figure 8.

X matrix can get three matrices $X=WSP^T$ after SVD operation. W is the term vector of X, S is the singular value matrix of X, P^T is the document vector of X. Like many papers (Deerwester et al., 1990; Landauer et al., 1998), for the singular value matrix, the top 2 singular value is selected. Then reestablishes the three matrixes can obtain new matrix $X'=W'S'P'^T$, that is LSA matrix. The LSA can strengthen the latent semantic characteristic of keywords. While comparing matrix of Figure 8 and LSA matrix of Figure 9, we can find TF-IDF value of every vocabulary is nearly 0 in original X matrix but the value of LSA matrix obviously increases after dimension reduction and rebuild matrix. That expresses the LSA has found the semantic dependence of each keyword and basketball.

X Matrix.csv

	A	B	C	D	E	F	G	H	I	J	K	L	M
7	失誤	0	0	0	0	0	0	0	0	0	0	0	0
8	教練團	0	0	0	0	0	0	0	0	0	0	0	0
9	戰將	0	0.157141	0	0	0.392852	0	0	0	0	0	0	0
10	開幕戰	0	0	0	0	0	0	0	0	0	0	0	0
11	資格賽	0	0	0	0	0	0	0	0	0	0	0	0
12	冠軍	0	0	0	0	0	0	0	0	0	0	0	0
13	總冠軍	0	0	0	0	0	0	0	0	0	0	0	0
14	季後賽	0	0	0	0	0	0	0	0	0	0	0	0
15	三分	0	0	0	0	0	0	0	0	0	0	0	0
16	表演賽	0	0	0	0	0	0	0	0	0	0	0	0
17	決賽	0	0	0	0	0	0	0	0	0	0	0	0
18	選秀會	0	0	0	0	0	0	0	0	0	0	0	0
19	獎盃	0	0	0.1404	0	0.386101	0	0	0	0	0	0	0
20	主場	0	0	0	0	0	0	0	0	0	0	0	0
21	職籃	0	0	0	0	0	0	0	0	0	0	0	0
22	選秀權	0	0	0	0	0	0	0	0	0	0	0	0
23	新人王	0	0	0	0	0	0	0	0	0	0	0	0

X Matrix

Figure 8. Terms-documents matrix: X.

Finally, the system transform LSA matrix of X to the high dimensional space. Utilize Euclidean distance function to calculate the distance between the new keyword and concepts of the ontology. The new keyword and the ontology concept distance is near expressed both semantic are close. Table 1 expresses the shortest distance between new keywords and the concepts of the ontology. The value of distance is smaller, the semantic is more similar. While the distances between new keyword and concepts have many various values, the system will choose the one with minimum value of distance. Consequently, we select the "***總決賽***", "***控衛***", "***獎盃***", "***戰將***" and "***籃賽***" five new keywords to insert the ontology. The Figure 10 is show the new basketball ontology.

LSA Matrix.csv

	A	B	C	D	E	F	G	H	I	J	K	L	M
7	失誤	1.78E-20	5.02E-20	1.96E-20	9.66E-21	3.55E-20	6.43E-21	3.72E-20	6.51E-21	1.34E-21	3.69E-21	2.66E-22	2.92E-21
8	教練團	9.52E-20	2.69E-19	1.05E-19	5.17E-20	1.90E-19	3.44E-20	1.99E-19	3.49E-20	7.18E-21	1.97E-20	1.42E-21	1.56E-20
9	戰將	0.031177	0.086323	0.03568	0.017616	0.061209	0.01127	0.066143	0.011437	0.002334	0.006725	4.59E-04	0.005092
10	開幕戰	3.25E-20	9.17E-20	3.58E-20	1.76E-20	6.48E-20	1.17E-20	6.80E-20	1.19E-20	2.45E-21	6.74E-21	4.85E-22	5.33E-21
11	資格賽	1.46E-20	4.12E-20	1.61E-20	7.93E-21	2.91E-20	5.27E-21	3.05E-20	5.34E-21	1.10E-21	3.03E-21	2.18E-22	2.40E-21
12	冠軍	-6.08E-22	-1.72E-21	-6.70E-22	-3.30E-22	-1.21E-21	-2.20E-22	-1.27E-21	-2.23E-22	-4.59E-23	-1.26E-22	-9.08E-24	-9.99E-23
13	總冠軍	1.58E-22	4.47E-22	1.74E-22	8.60E-23	3.16E-22	5.73E-23	3.32E-22	5.80E-23	1.20E-23	3.29E-23	2.37E-24	2.60E-23
14	季後賽	-2.52E-23	-7.12E-23	-2.78E-23	-1.37E-23	-5.04E-23	-9.12E-24	-5.28E-23	-9.24E-24	-1.90E-24	-5.23E-24	-3.77E-25	-4.15E-24
15	三分	-2.18E-24	-6.16E-24	-2.40E-24	-1.19E-24	-4.35E-24	-7.89E-25	-4.57E-24	-7.99E-25	-1.65E-25	-4.53E-25	-3.26E-26	-3.58E-25
16	表演賽	-2.27E-25	-6.42E-25	-2.50E-25	-1.24E-25	-4.54E-25	-8.22E-26	-4.76E-25	-8.33E-26	-1.72E-26	-4.72E-26	-3.40E-27	-3.74E-26
17	決賽	5.18E-26	1.46E-25	5.70E-26	2.81E-26	1.03E-25	1.87E-26	1.08E-25	1.90E-26	3.91E-27	1.07E-26	7.73E-28	8.51E-27
18	選秀會	1.30E-27	3.66E-27	1.43E-27	7.05E-28	2.59E-27	4.69E-28	2.72E-27	4.75E-28	9.79E-29	2.69E-28	1.94E-29	2.13E-28
19	獎盃	0.015544	0.040776	0.019519	0.009653	0.029159	0.00561	0.034065	0.005724	0.001138	0.003684	2.19E-04	0.002497
20	主場	2.87E-29	8.10E-29	3.16E-29	1.56E-29	5.73E-29	1.04E-29	6.01E-29	1.05E-29	2.17E-30	5.95E-30	4.29E-31	4.72E-30
21	職籃	2.40E-31	6.78E-31	2.64E-31	1.30E-31	4.79E-31	8.68E-32	5.03E-31	8.79E-32	1.81E-32	4.98E-32	3.59E-33	3.94E-32
22	選秀權	1.31E-31	3.69E-31	1.44E-31	7.11E-32	2.61E-31	4.73E-32	2.74E-31	4.79E-32	9.87E-33	2.71E-32	1.95E-33	2.15E-32
23	新人王	-2.03E-33	-5.73E-33	-2.24E-33	-1.10E-33	-4.05E-33	-7.34E-34	-4.25E-33	-7.44E-34	-1.53E-34	-4.21E-34	-3.03E-35	-3.34E-34

LSA Matrix

Figure 9. LSA matrix of X.

Table 1. The distance between new keyword and ontology concept.

New	Ontology concept	Distance
總決賽	***NBA***	***0.138432***
戰績	NBA	0.334543
控衛	***快攻***	***0.280933***
獎盃	***抄截***	***0.155808***
戰將	***進攻***	***0.41454***
籃賽	***籃球***	***0.010934***
三連勝	籃球	0.041226
三分	籃球	0.044276
三分線	籃球	0.044276
主場	籃球	0.044276
主導權	籃球	0.044276
失誤	籃球	0.044276

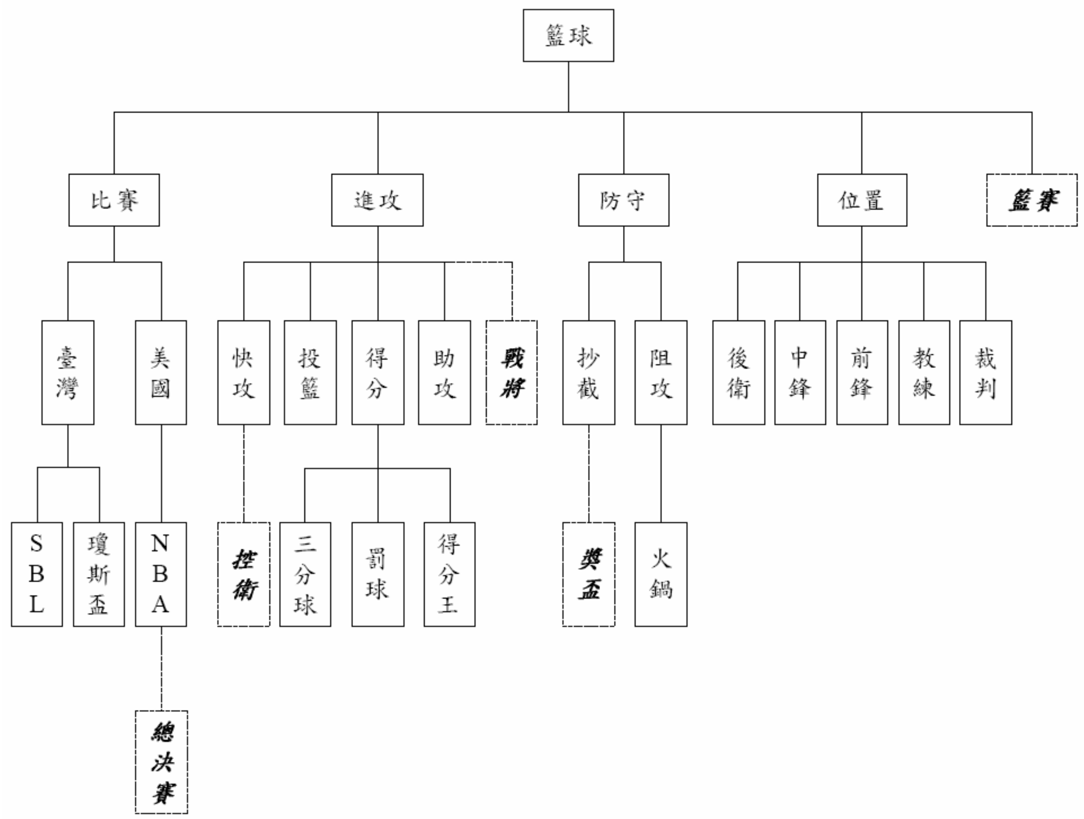

Figure 10. The new basketball ontology.

5. CONCLUSION

At present, Most of research is about the ontology construction but along with time changing with each new day, the already constructed the ontology possibly does not match the user requirement. Therefore, we try to propose a method of adding new concepts on the domain ontology. The system utilizes TF-IDF to find out the relevant and important keywords from the web pages. Through LSA semantic analyses to find the keywords that have highest similarity between the concepts of ontology. Then the new keyword insert to the domain ontology. This method can be fast renewal ontology, but does not spend a large amount of manpower and time to constructs and maintenance ontology. According to experimental result, that shows the method of us proposes can really be effective renewal ontology but still need some methods to evaluate the effect even add some semi-automatic methods to modify the results. In the future, we plane to make an ontology generating system accord with the user's need and accelerate the forming of the semantic web.

REFERENCES

Abasolo, J. & Gómez, M. 2000. MELISA. An ontology-based agent for information retrieval in medicine. *ECDL 2000 Workshop on the Semantic Web Lisbon*, 73-82.

Berners-Lee, T., Fischetti, M. & Dertouzos, M. 1999. Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by Its Inventor. *Harper San Francisco*, 1st edition.

Berners-Lee, T., Hendler, J. & Lassila, O. 2001. The Semantic Web. *Scientific American*, 34-43.

Broekstra, J., Klein, M., Decker, S., Fensel, D., van Harmelen, F. & Horrocks, I. 2002. Enabling knowledge representation on the Web by extending RDF Schema. *Computer Networks*, 39(5), 609-634.

Buggenhout, C. V. & Ceusters, W. 2005. A novel view on information content of concepts in a large ontology and a view on the structure and the quality of the ontology. *International Journal of Medical Informatics*, 74(2-4), 125-132.

Chen, L., Tokuda, N. & Nagai, A. 2003. A new differential LSI space-based probabilistic document classifier. *Information Processing Letters*, 88(5), 203-212.

Chen, R. C., Liang, J. Y. & Pan, R. H. 2005. Constructing an Ontology based on ART Neural Network. *Proceedings of the 10th Artificial Intelligence and Applications (TAAI-2005)*, Kaohsiung.

Chiang, R. H. L., Cecil, C. E. H. & Storey, V. C. 2001. A smart Web query method for semantic retrieval of Web data. *Data & Knowledge Engineering*, 38(1), 63-84.

CKIP on-line Chinese words segmented service. *Chinese Knowledge Information Processing Group of Academia Sinica* [On-line], Available: http://ckipsvr.iis.sinica.edu.tw/, Access time 2005/09/29.

Deerwester, S., Dumais, S. T., Landauer, T. K., Furnas, G. W. & Harshman, R. 1990. Indexing by Latent Semantic Analysis. *Journal of the American Society of Information Science*, 41(6), 391-407.

Gruber, T. 1993. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2), 199-220.

Guarino, N. 1998. Formal Ontology and Information System. *In Proceedings of the Formal Ontology and Information System (FOIS'98)*, Trento, Italy, 3-15.

Han, L., Sun, L., Chen, G. & Xie, L. 2006. ADSS: An approach to determining semantic similarity. *Advances in Engineering Software*, 37(2), 129-132.

Jacob, E. K. 2003. Ontologies and the Semantic Web. *Bulletin of the American Society for Information Science and Technology*, 19-22.

Jenkins, C., Jackson, M., Burden, P. & Wallis, J. 1999. Automatic RDF metadata generation for resource discovery. *Proceeding of the eighth international conference on World Wide Web*, Toronto, Canada, 1305-1320.

Khan, L. & Luo, F. 2002. Ontology Construction for Information Selection. *14th IEEE International Conference on Tools with Artificial Intelligence(ICTAI'02)*, 122-127.

Landauer, T. K., Foltz, P. W. & Laham, D. 1998. An introduction to Latent Semantic Analysis. *Discourse Processes*, 25, 259-284.

Maedche, A. & Staab, S. 2000. Discovering Conceptual Relations from Text. in W. Horn, ed., *'Proceedings of the 14th European Conference on Artificial Intelligence(ECAI 2000)'*, IOS Press, Berlin, Germany, 321–325.

Manola, F. & Miller, E. 2004. RDF Primer. *World Wide Web Consortium* [On-line], Available: http://www.w3.org/TR/rdf-primer/, Access time 2004/10/15.

Meziane, F. & Rezgui, Y. 2004. A document management methodology based on similarity contents.*Information Sciences—Informatics and Computer Science: An International Journal*, 158(1), 15-36.

Miller, E. 2004. Weaving Meaning : An Overview of The Semantic Web. *World Wide Web Consortium* [On-line], Available: http://www.w3.org/2004/Talks/0120-semweb-umich/slide8-0.html, Access time 2005/12/15.

Oyama, S., Kokubo, T. & Ishida, T. 2004. Domain-Specific Web Search with Keyword Spices. *IEEE Transactions on knowledge and data engineering*, 16(1), 17-27.

Patel, C., Supekar, K., Lee, Y. & Park, E. K. 2003. OntoKhoj: a semantic web portal for ontology searching, ranking and classification. *Proceedings of the 5th ACM international workshop on Web information and data management(WIDM 2003),* 58-61.

Salton, G. & McGill, M. J. 1983. Introduction to Modern Information Retrieval. *McGraw Hill Book Co.*.

Shen, J. & Yang, Y. 2004. Extending RDF in distributed knowledge-intensive applications. *Future Generation Computer Systems*, 20(1), 27-46.

Stefan, D., Frank van, H., Jeen, B., Michael, E., Dieter, F., Ian, H., Michel, K. & Sergey, M. 2000. The Semantic Web - on the respective Roles of XML and RDF. *IEEE Internet Computing,* 4(5), 63-73.

Stephens, L. M. & Huhns, M. N. 2001. Consensus Ontologies Reconciling the Semantics of Web Pages and Agents. *IEEE Internet Computing,* 5(5), 92-95.

Tan, K. W., Han H. & Elmasri, R. 2000. Web Data Cleansing and Preparation for Ontology Extraction using WordNet. *First International Conference on Web Information Systems Engineering(WISE'00)*, 2, 11-18.

Wei, J., Bressan S. & Ooi, B. C. 2000. Mining Term Association Rules for Automatic Global Query Expansion: Methodology and Preliminary Results. *In Proceedings of the First International Conference on Web Information Systems Engineering (WISE'00),* 366-373.

Williams, A. B. & Tsatsoulis, C. 2000. An Instance-based Approach for Identifying Candidate Ontology Relations within a Multi-Agent System. *Fourteenth European Conference on Artificial Intelligence, Ontology Learning ECAI-2000 Workshop,* Berlin, Germany.