UNIVERSITI SAINS MALAYSIA

Pejabat Pengurusan Dan Kreativiti Penyelidikan
Research Creativity and Management Office

Canselori,

Universiti Sains Malaysia
Aras 6, Bangunan Canselori
11800, USM Pulau Pinang
T : (6)04-653 3108/3178/3988/5019
F : (6)04-656 6466/8470
: (6)04-653 2350
L : www.research.usm.my

No. Fail : F0288
Tarikh : 2Disember 2011

Encik Gian Chand @ Sodhy A/L Didar Singh
Pusat Pengajian Sains Komputer
Universiti Sains Malaysia

Tuan,

LAPORAN AKHIR SKIM GERAN PENYELIDIKAN FUNDAMENTAL (FRGS)
Tajuk Projek : Semantic-based Document Categorization using Ontology
No. Akaun : 203/PKOMP/671178

Dengan hormatnya perkara di atas dirujuk.

2. Terlebih dahulu saya ucapkan ribuan terima kasih di atas satu salinan laporan akhir untuk projek penyelidikan seperti tajuk di atas.

3. Adalah dimaklumkan walaupun projek ini telah selesai, kerjasama Jabatan Bendahari dipohon untuk menguruskan penutupan akaun projek pada selewat-lewatnya **31 Disember 2011**. Tempoh ini bertujuan untuk menyelesaikan semua urusan tuntutan dan bayaran yang telah dibelanjakan di dalam tempoh projek. Walau bagaimanapun, tuan dinasihatkan supaya tidak mengeluarkan borang-borang pesanan baru di dalam tempoh ini.

4. Selanjutnya sila ambil perhatian terhadap perkara-perkara berikut sekiranya berkaitan:

(i) Semua penerbitan harus merakamkan penghargaan kepada **Skim Geran Penyelidikan Fundamental (FRGS)** dan tuan dipohon mengemukakan satu salinan ke Pejabat ini.

(ii) Bahagian Penyelidikan & Inovasi boleh/akan mengagihkan semula peralatan yang telah dibeli menggunakan peruntukan geran ini seandainya terdapat penyelidik lain yang memerlukan peralatan tersebut.

5. Akhir sekali, tahniah di atas usaha dan kejayaan pihak tuan dapat menyelesaikan projek ini dengan jayanya.

Sekian, terima kasih.

"BERKHIDMAT UNTUK NEGARA"
'Memastikan Kelestarian Hari Esok'

Yang menjalankan tugas,

(AMRA OTHMAN)
Penolong Pendaftar
Unit Pengurusan Geran & Kontrak

HAH, HAR, SM

LAPORAN AKHIR SKIM GERAN PENYELIDIKAN FUNDAMENTAL (FRGS)
Tajuk Projek : Semantic-based Document Categorization using Ontology
No. Akaun : 203/PKOMP/671178

s.k. Dekan Penyelidikan
Pelantar Sains Fundamental
Pejabat Pelantar Penyelidikan
Universiti Sains Malaysia

Dekan
Pusat Pengajian Sains Komputer
Universiti Sains Malaysia

Timbalan Dekan
(Pengajian Siswazah & Penyelidikan)
Pusat Pengajian Sains Komputer
Universiti Sains Malaysia

Ketua Pustakawan
Perpustakaan Hamzah Sendut
Universiti Sains Malaysia

} Disampaikan satu salinan laporan akhir projek untuk simpanan Perpustakaan

Penolong Bendahari Kanan
Unit Kumpulan Wang Penyelidikan
Jabatan Bendahari
Universiti Sains Malaysia

Pegawai Sains
Pelantar Sains Fundamental
Pejabat Pelantar Penyelidikan
Universiti Sains Malaysia

} Mohon kerjasama pihak puan untuk menguruskan penutupan akaun projek selewat-lewatnya pada **31 Disember 2011** dan mohon kemukakan satu salinan penyata kewangan terakhir ke Pejabat ini untuk tujuan rekod

BORANG FRGS – P3(R)

FINAL REPORT
FUNDAMENTAL RESEARCH GRANT SCHEME (FRGS)

Laporan Akhir Skim Geran Penyelidikan Asas (FRGS) IPT
Pindaan 1/2010

A

RESEARCH TITLE : **SEMANTIC-BASED DOCUMENT CATEGORIZATION USING ONTOLOGY**
Tajuk Penyelidikan

PROJECT LEADER : **EN. GIAN CHAND SODHY**
Ketua Projek

PROJECT MEMBERS : 1. **PROF. MADYA DR. TANG ENYA KONG**
(including GRA) 2. **DR. RANAIVO-MALANÇON BALISOAMANANDRAY**
Ahli Projek 3. **CIK. SARAVADEE SAE TAN**

PROJECT ACHIEVEMENT *(Prestasi Projek)*

B

ACHIEVEMENT PERCENTAGE			
Project progress according to milestones achieved up to this period	**0 - 50%**	**51 - 75%**	**76 - 100%**
Percentage			100%

RESEARCH OUTPUT		
Number of articles/ manuscripts/ books *(Please attach the First Page of Publication)*	**Indexed Journal**	***Non-Indexed Journal***
Conference Proceeding *(Please attach the First Page of Publication)*	**International**	***National***
	1	
Intellectual Property *(Please specify)*		

HUMAN CAPITAL DEVELOPMENT

Human Capital	**Number**				**Others** (please specify)
	On-going		**Graduated**		
Citizen	Malaysian	Non Malaysian	Malaysian	Non Malaysian	
PhD Student	1				
Master Student					
Undergraduate Student					
Total	1				

EXPENDITURE *(Perbelanjaan)*

C

Budget Approved *(Peruntukan diluluskan)* : **RM 40,000.00**
Amount Spent *(Jumlah Perbelanjaan)* : **RM 39,967.59**

Balance *(Baki)* : **RM 32.41**
Percentage of Amount Spent : **99.92 %**
(Peratusan Belanja)

ADDITIONAL RESEARCH ACTIVITIES THAT CONTRIBUTE TOWARDS DEVELOPING SOFT AND HARD SKILLS
(Aktiviti Penyelidikan Sampingan yang menyumbang kepada pembangunan kemahiran insaniah)

D

International		
Activity	Date (Month, Year)	Organizer
(e.g : Course/ Seminar/ Symposium/ Conference/ Workshop/ Site Visit)		

National		
Activity	Date (Month, Year)	Organizer
(e.g : Course/ Seminar/ Symposium/ Conference/ Workshop/ Site Visit)		

PROBLEMS / CONSTRAINTS IF ANY *(Masalah/ Kekangan sekiranya ada)*

E

RECOMMENDATION (*Cadangan Penambahbaikan*)

F

RESEARCH ABSTRACT – Not More Than 200 Words (*Abstrak Penyelidikan – Tidak Melebihi 200 patah perkataan*)

G

Text categorization, the assignment of free text documents to one or more predefined categories based on their content, is an important component in many information management task. We have observed that the knowledge represented in the category model has a significant impact for the success of a text categorization task. In this research, we use Wikipedia as a knowledge source to enrich our category model.

Wikipedia is a well known, valuable and free information repository on the Web, constructed in a collaborative effort by voluntary contributors. In Wikipedia, each article describes a specific entity in the world and has a unique name in the collection. It is a rich and useful source of knowledge. Wikipedia categories, Wikipedia link structure and Infoboxes are valuable information and can be leveraged in many applications such as text categorization. In our research, we selected 8 categories from Wikipedia (i.e. 'Film', 'Actor', 'Singer', 'Music group'. 'Film Producer'. 'Film Director', 'Album', and 'Song') and crawled 445 articles randomly from these categories as our training data. We used a feature selection method that learns the conceptual information and descriptive information from the Wikipedia articles by utilizing the heading text, infobox and hyperlink of the articles. Our studies show promising results that Wikipedia is an important knowledge source and can be used to enrich our category model.

Publication:

- Saravadee Sae Tan, Tang Enya Kong and Gian Chand Sodhy, **Annotating Wikipedia Articles with Semantic Tags for Structured Retrieval**, Proceedings of the *2nd Workshop on Social Web Search and Mining (SWSM2009)*, Hong Kong, 2-6 November 2009, pp. 17-24.

Date : 5 March 2010
Tarikh

Project Leader's Signature:
Tandatangan Ketua Projek

COMMENTS, IF ANY/ ENDORSEMENT BY RESEARCH MANAGEMENT CENTER (RMC)
(Komen, sekiranya ada/ Pengesahan oleh Pusat Pengurusan Penyelidikan)

H .....

.....

.....

Name:
Nama:

Signature:
Tandatangan:

Date:
Tarikh:

Annotating Wikipedia Articles with Semantic Tags for Structured Retrieval

Saravadee Sae Tan
Faculty of Information Technology
Multimedia University,
Selangor, Malaysia
+6 (019) 520 4161
saratan@cs.usm.my

Tang Enya Kong
Faculty of Information Technology
Multimedia University,
Selangor, Malaysia
+6 (03) 8312 5054
enyakong@mmu.edu.my

Gian Chand Sodhy
School of Computer Sciences
Universiti Sains Malaysia
Penang, Malaysia
+6 (04) 653 3002
sodhy@cs.usm.my

ABSTRACT

Structured retrieval aims at exploiting the structural information of documents when searching for documents. Structured retrieval makes use of both content and structure of documents to improve information retrieval. Therefore, the availability of semantic structure in the documents is an important factor for the success of structured retrieval. However, the majority of documents in the Web still lack semantically-rich structure. This motivates us to automatically identify the semantic information in web documents and explicitly annotate the information with semantic tags.

Based on the well-known Wikipedia corpus, this paper describes an unsupervised learning approach to identify conceptual information and descriptive information of an entity described in a Wikipedia article. Our approach utilizes Wikipedia link structure and Infobox information in order to learn the semantic structure of the Wikipedia articles. We also describe a lazy approach used in the learning process. By utilizing the Wikipedia categories provided by the contributors, only a subset of entities in a Wikipedia category is used as training data in the learning process and the results can be applied to the rest of the entities in the category.

Categories and Subject Descriptors

H.3.3 **[Information Storage and Retrieval]**: Information Search and Retrieval – *retrieval models, search process.*

General Terms

Algorithms, Design.

Keywords

Structured Retrieval, Semantic Markup, Wikipedia,

SWSM'09, November 2, 2009, Hong Kong, China.

1. INTRODUCTION

Extensible Markup Language or XML [1] has become the most widely used standard for marking up the content of a document. The intention of adding XML markup to a document is to make the structure of the document explicit. XML markup adds extra information to a document and allows computer systems to easily understand and search the content of the document. Two main types of structure in XML documents are logical structure and semantic structure [11]. Logical structure represents the logical organization of the document components. It commonly describes the division of a document, such as a document is divided into chapters and chapters into sections. On the other hand, semantic structure gives semantic information about the content. For example, we can indicate that a phrase is an author name or a text fragment is an address of a restaurant.

Structured retrieval, also referred to as XML retrieval, aims at exploiting the structural information of a document in order to implement a more focused retrieval strategy and return document fragments instead of whole documents in response to a user query [3]. By taking advantage of the structural information, the retrieval is improved in two aspects: increased functionality and increased precision [7][8]. Here, we describe several ways in which the semantic structure in a document can be leveraged to improve information retrieval.

- **Specifying Semantic Constraints in Query.** Traditional information retrieval allows users to express their information needs through a query, which is usually composed of a list of keywords. Keyword-based queries are simple and easy to use, but they only specify the content constraints. For example, the query "Steven Seagal actor" only specifies that the documents returned should contain these keywords. Structured retrieval allows the specification of semantic constraints, in addition to content constraint, in a query. Semantic constraints enable a user to describe his information needs more precisely. For example, if a user is looking for information about Steven Seagal's actor career, he can specify structural query <actor>Steven Seagal</actor> to restrict the search to documents that containing Steven Seagal as an actor.
- **Disambiguating Named Entity.** A term may have multiple senses or meanings in a corpus and precision suffers if the information retrieval system is unable to distinguish the meanings of the terms in a document. The semantic markup in a document can clearly indicate the meaning of a term. This allows the information retrieval system to restrict its search to

a particular meaning of the term. For example, the term "starry night" may refer to the song written by Don Mclean, or the live album by Don Mclean, or the piece of painting by Vincent van Gogh. If a user's intention is to find information about the song, he can restrict his search to documents in which the term "starry night" is annotated with the concept <song>.

- **Defining the Role of Named Entity.** A specific entity may have different roles in different contexts. Besides disambiguating entities with same surface text, there is also a need to explicitly specify the roles of an entity in different contexts. For instance, Steven Seagal is not only known as a famous actor, director and producer, but also as a guitarist and singer. The semantic markup can be used to annotate the roles of an entity such as <actor>, <director> and <producer> in the documents. Consider the query "find movie directed by Steven Seagal". By exploiting the semantic markup, the information retrieval system is able to retrieve only documents in which the term "Steven Seagal" is annotated with <director>. Therefore, only movies directed by Steven Seagal are returned as results, whereas movies starring by Steven Seagal but not directed by him are eliminated.

The availability of semantic structure in documents is an important factor for the success of structured retrieval. However, despite the advantages and simplicity of XML, the majority of documents on the Web still lack semantic structures. This is because manually marking up the semantic structures of documents is tedious and expensive, whereas automating semantic mark up of documents is still a challenging task.

Based on the well-known Wikipedia corpus, this paper describes an approach to mark up Wikipedia articles with semantic tags that describe the concept of the entities in the articles. In our approach, we exploit three types of information from Wikipedia corpus: Wikipedia categories associated to the articles, Wikipedia link structure and Infoboxes in Wikipedia articles.

2. WIKIPEDIA CORPUS

Wikipedia is a well known, valuable and free information repository on the Web, constructed in a collaborative effort by voluntary contributors. In Wikipedia, each article describes a specific entity in the World and has a unique name in the collection. For example, the page *Steven_seagal* describes the American actor Steven F. Seagal, the page *Under_Siege* describes the action movie Under Seige and the page *Songs_from_the_Crystal_Cave* describes the music album by Steven Seagal.

Wikipedia articles are highly interconnected via hyperlinks. In a Wikipedia article, most of the entities that exist in the content are linked to the articles describing the entities. The link structure of Wikipedia articles provides valuable information about the relationship between entities in the collection.

Most of the Wikipedia articles are associated with one or more categories by the contributors. The page *Steven_seagal* is associated to the categories *American_film_actors, American_film_directors, American_male_singers, 1951_births, American_aikidoka,* and 14 more. Each Wikipedia category represents a particular concept and the entity that belongs to the category can be considered as an instance of the concept. For example, articles associated with the category *American_film_actors* are describing those film actors whose nationality is American. Detail description of Wikipedia categories structure is discussed in [9][10].

Many Wikipedia articles contain infoboxes that give a general description of the entities described in the articles. Infobox is a summary of the entity's most representative attributes organized in a table. For example, the page *Steven_seagal* contains an Infobox describing the Steven Seagal's date and place of birth, other names and also his spouse, whereas the Infobox in the page *Under_Siege* lists the director, producer, actors, screenwriters and other information about the movie.

Although wikipedia articles are rich source of information, they are mainly for human reading and not for machine to process. Therefore, there is a need to explicitly mark up the Wikipedia articles with semantic tags to allow a computer system to easily understand and search the content. Apart from text content, Wikipedia articles contain various extra information such as Wikipedia categories, Wikipedia link structure and Infoboxes added by contributors collaboratively. This information is valuable and can be leveraged to learn the semantic structure of the Wikipedia articles.

3. SEMANTIC MARKUP OF WIKIPEDIA ARTICLE

Inevitably, the semantic structure in a document is important to improve information retrieval. The semantic structure can be exploited by the information retrieval system in order to retrieve only those results that satisfy a user's information need. Rather than relying on the contributors to add semantic markup when creating Wikipedia articles, we propose to automatically identify the semantic information we intend to leverage for structured retrieval. The Wikipedia categories are used as the basis for our approach.

In Wikipedia corpus, most Wikipedia articles are associated to a list of categories. The categories are assigned to the articles by Wikipedia contributors in a collaborative manner. Therefore, the categories can be considered as important description of the entity mentioned in a Wikipedia article. Figure 1 below gives an example of categories associated with the page *Steven_Seagal* (http://en.wikipedia.org/wiki/Steven_Seagal).

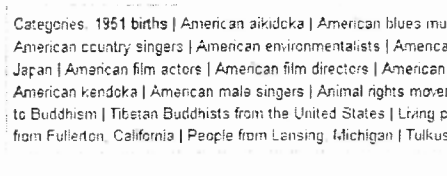
Categories: 1951 births | American aikidoka | American blues musicians | American country singers | American environmentalists | American expatriates in Japan | American film actors | American film directors | American guitarists | American kendoka | American male singers | Animal rights movement | Converts to Buddhism | Tibetan Buddhists from the United States | Living people | People from Fullerton, California | People from Lansing, Michigan | Tulkus | Rinpoches

Figure 1: List of Wikipedia categories associated to *Steven_Seagal* page

We found that Wikipedia categories can provide two types of information to describe the entities associated with them. In this paper, we refer to these two types of information as conceptual information and descriptive information. Some of the Wikipedia categories describe the concept or class of an entity, such as the category *Guitarists* implies that the person described in an article

is a musician who plays the guitar. On the other hand, some categories provide information about the attributes of an entity. For example, the category *1951_births* indicates the year the person was born, whereas the category *People_from_Lansing,_Michigan* indicates the person is the resident of the city Lansing, Michigan. There are some categories that provide both conceptual information and descriptive information, for instance, the category *American_guitarists* implies the class of the entity (a guitarist) and also the entity's attribute (he is from the United State).

In this paper, we describe an unsupervised learning approach to identify the conceptual information and descriptive information from the Wikipedia articles and explicitly mark up the information with semantic tags. Note that in this paper we will use the term 'Wikipedia categories' to refer to the categories assigned to Wikipedia articles by the contributors to avoid confusion with the category (or concept) an entity naturally belong to.

3.1 Learning Conceptual Information

In Wikipedia corpus, there are a large number of named entities and new entities appear every day in the collection. Without any prior knowledge about the possible concepts described by the entities in the corpus, we propose an unsupervised learning algorithm to identify the concepts described in Wikipedia. The task of identifying concepts from Wikipedia can be formulated as a categorization problem. Entities of a particular concept will have a set of similar attributes. For example, 'Under Seige' and 'On Deadly Ground' are two different entities but belong to the same concept, 'film'. Both entities have a similar set of attributes, such as a director who direct the making of the movie, the producer of the movie, the cast, and etc. Therefore, entities of a same concept can be identified by measuring their common attributes.

3.1.1 Named Entity Categorization

The categorization approach consists of three main steps:

3.1.1.1 Feature Selection

In Feature Selection process, a set of features is extracted from the Wikipedia article to represent the entity mentioned in the article. Features are keywords or descriptive terms in a text that can represent a category. In our approach, two types of terms are extracted as features:

Heading Text. In Wikipedia articles, heading text usually describes the topic of the content below the header. For example, in the page *Under_Seige* (http://en.wikipedia.org/wiki/Under_Seige) (Figure 2), the content below the header 'Plot' describes the main story of the movie, whereas the header 'Cast' describes a list of actors along with their roles in the movie. Hence, we extract all heading text from a Wikipedia article as features to represent the entity.

Infobox Parameter Name. Wikipedia Infobox provides summary of an entity's key attributes. In Figure 2, Infobox in the page *Under_Seige* gives the standard information about a movie, such as the director, producer, actors and etc. The attributes are represented by parameter names (e.g. directed by, produced by) and parameter values (e.g. Andrew Davis). As the parameter name provides a clear indication of the semantics of the attribute, we also extract all parameter names as our features.

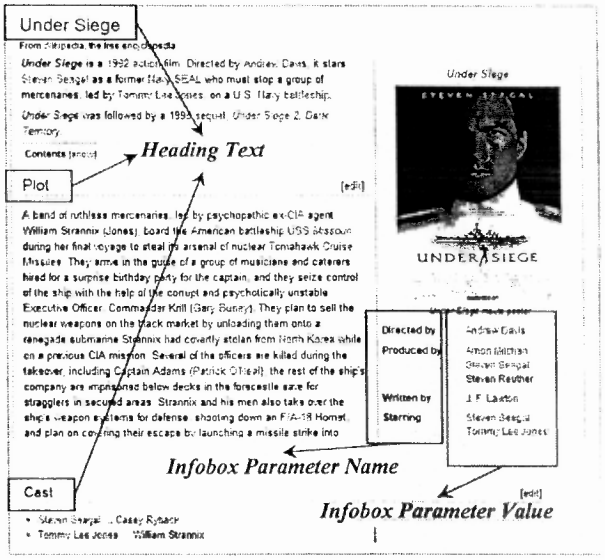

Figure 2: Feature Snapshot of an entity in Wikipedia

3.1.1.2 Similarity Measure

The entities and their features are represented in a Vector Space Model [5]. Each entity e is represented as a vector v and each dimension in the vector corresponds to a feature. If a feature occurs in the article, the value is set to 1; else the value is set to 0. The similarity between two entities is measured by calculating the cosine value of the vectors that represent the entities.

$$sim(e_1,e_2) = \cos\theta(v_1,v_2) = \frac{v_1 \cdot v_2}{\|v_1\|\|v_2\|}$$

3.1.1.3 Entity Clustering

An agglomerative clustering method is used to categorize entities that belong to a same concept or category. Given X is the set of entities $X=\{e_1,e_2,\ldots,e_k\}$ and k is the total number of entities in the corpus, we calculate similarity for each of the $\frac{k(k-1)}{2}$ pairs of entities in X. For each entity e_i, we find the entity e_j with maximum similarity with e_i.

$$sim(e_i,e_j) = \arg\max_{e_k \in X} sim(e_i,e_k)$$

If the similarity value between two entities is greater than a predefined threshold, τ, we assume that both entities belong to a same concept and are categorized together.

$$categorize(e_i,e_j) = \begin{cases} true & \text{if } sim(e_i,e_j) \geq \tau \\ false & \text{otherwise} \end{cases}$$

3.1.2 Category Relation Learning

Some entities naturally fall into more than one category or concept. An entity may refer to a different concept in a different context. For example, Steven Seagal is referred as a 'singer' in

documents discussing his music, whereas he is referred as an 'actor' in documents discussing his movie. In order to identify the appropriate concept of an entity in different contexts, we learn the relations between concepts in Wikipedia corpus.

We define relation between two concepts $R_{c1,\ c2}$ as follow: Given X is the set of entity in concept c_1 and Y is the set of entity in concept c_2. Concept c_1 is related to concept c_2 if there is an entity $e_i \in X$ that link to another entity $e_j \in Y$. Figure 3 illustrate the relations between two Wikipedia categories.

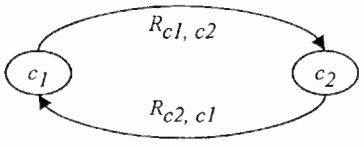

Figure 3: Illustration of Relations between two Wikipedia Categories

The weight of a relation $R_{c1,c2}$ is calculated based on two factors: the link frequency (*lf*) and the entity frequency (*ef*). Link frequency is the total number of outgoing links from X to Y, whereas entity frequency is the number of entity in X that contains at least one link to Y.

$$weight(R_{c1,c2}) = \frac{lf}{N_l} + \frac{ef}{N_e}$$

We consider the relative frequency by normalizing *lf* and *ef*. N_l is total number of links and N_e is total number of entities in the collection.

3.1.3 Extension: A Lazy Approach

The number of entities in Wikipedia is large and this number increases as new articles is added everyday. Hence, calculating the similarity between entities becomes a time consuming and costly task. Here, we describe an effective way to categorize named entities by utilizing the Wikipedia categories in the articles. We make an assumption that all entities associated to a Wikipedia category belong to a same concept. Therefore, instead of calculating the similarity between entities, we calculate similarity between Wikipedia categories. Our aim here is to identify Wikipedia categories that are referring to a same concept.

For each Wikipedia category, only a subset of the articles is used as training data for the learning process. Each Wikipedia category is represented by a set of features extracted from the training data. Similarity measure between two Wikipedia categories is calculated and if the value is greater than a threshold, all entities of the two Wikipedia categories are considered as belonging to a same concept.

3.2 Learning Descriptive Information

The task of identifying descriptive attributes of the entities in a Wikipedia category can be formulated as a feature weighting problem. The main aim here is to identify descriptive features from the articles in a Wikipedia category.

3.2.1 Descriptive Feature Selection

3.2.1.1 Feature Candidature

A hyperlink in a Wikipedia article indicates that the entity relates to another entity (the target Wikipedia article the link points to) in some respect. The target entity usually provides further description of the source entity. For a Wikipedia category, all links in the training data are extracted as candidate features in our approach.

3.2.1.2 Feature Weighting

We calculate a weight for each candidate feature in order to identify descriptive features. In this paper, we use two indicators to weight a feature.

Frequency Occurrence. A feature is considered as representative for a Wikipedia category c if it appears in many articles in c. Here, we calculate the entity frequency (*ef*) for each candidate feature f_i in c. Entity frequency implies the number of entities (Wikipedia articles) in which f_i occurs at least once. We consider the relative frequency by normalizing *ef* with total number of entities N_e in c.

$$w1(f_i, c) = \frac{ef(f_i, c)}{N_e(c)}$$

Infobox Parameter Value. Wikipedia Infobox provides a set of fine-grained attributes to describe an entity. Each attribute in the Infobox demonstrates a 'is-a' relation between the parameter name and value, in which the parameter name is usually used to describe the concept of the entity mentioned in the parameter value. For example, in Figure 2, "Andrew Davis" is a "director" of a movie. In our approach, a candidate feature is given higher weight if it appears as Infobox parameter value. Given P_{value} is the set of parameter values extracted from all articles in a Wikipedia category, we compute a score for the feature f_i as follow:

$$w2(f_i, c) = \begin{cases} 1 & \text{if } f_i \in P_{value} \\ 0 & \text{otherwise} \end{cases}$$

By combining both *w1* and *w2*, we obtain a final weight for a candidate feature f_i in a Wikipedia category as follow:

$$weight(f_i, c) = \frac{1}{2}\left[w1(f_i, c) + w2(f_i, c)\right]$$

3.3 Experiments

3.3.1 Dataset

In our experiments, we define eight concepts that commonly occur in Wikipedia corpus, namely 'film, 'actor', 'singer', 'music group', 'film_producer', 'film_director', 'album', and 'song'. We collect 35 Wikipedia categories that belong to the defined concepts and crawl 445 articles randomly from the Wikipedia categories. Table 1 shows the dataset collected.

Table 1. The Dataset

Concept	Wikipedia Category
Film	American_films, 1990s_action_films 1990s_science_fiction_films, 1980s_action_films 1980s_science_fiction_films, Science_fiction_action_films
Actor	Actors_from_New_York, American_stage_actors Actors_from_New_Jersey, American_film_actors
Singer	English_singer-songwriters, English_male_singers Scottish_male_singers, American_rock_singers American_male_singers
Music Group	1960s_music_groups, 1980s_music_groups 1970s_music_groups
Film Producer	American_film_producers, English_film_producers Scottish_film_producers
Film Director	American_film_directors, English_film_directors Scottish_film_directors
Album	1990s_rock_album_stubs, 1980s_rock_album_stubs 1970s_rock_album_stubs, Don_McLean_albums Steven_Seagal_albums, Will_Smith_albums Elton_John_albums
Song	1980s_pop_songs, 1970s_pop_songs 1970s_single_stubs, Elton_John_songs

3.3.2 *Evaluation Measure*

As mentioned in Section 4.1.3, the main aim of our Named Entity Categorization problem is to assign two Wikipedia categories to the same cluster if they belong to a same concept. Therefore, the results are evaluated based on the commonly used measure in clustering algorithm.

Given N Wikipedia categories $C=\{c_1,c_2,\ldots,c_N\}$, we will have $\frac{N(N-1)}{2}$ pairs of Wikipedia categories, $<c_i, c_j>$. In the case where c_i and c_j belongs to a same concept, a true positive (*TP*) decision assigns c_i and c_j to the same cluster, whereas a false negative (*FN*) assigns c_i and c_j to different clusters. On the other hand, if c_i and c_j do not belong to the same concept, a true negative (*TN*) decision assigns c_i and c_j to different clusters, whereas a false positive (*FP*) assigns c_i and c_j to the same cluster. These notions are illustrated in the contingency table below (Table 2):

Table 2. The Illustration of TP, FN, TN and FP measure

	Belong to same concept	Belong to different concept
Categorized	true positive (*TP*)	false positive (*FP*)
Not categorized	false negative (*FN*)	true negative (*TN*)

Rand Index (*RI*) measures the percentage of decisions that are correct. *RI* ranges from 0 to 1. The value 1 indicates a perfect clustering, yet the value 0 indicates a total wrong clustering. *RI* is defined as below:

$$RI = \frac{TP + TN}{TP + FP + FN + TN}$$

Precision is the number of Wikipedia categories pairs $<c_i, c_j>$ correctly categorized (true positives) divided by the total number of pairs categorized (i.e. the sum of true positives and false positives). The *Precision* value 1 means that all pairs that belong to a same concept are successfully categorized.

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall is the number of correctly categorized pairs (true positives) divided by the total number of pairs that actually belong to a same concept (the sum of true positives and false negatives).

$$\text{Recall} = \frac{TP}{TP + FN}$$

Another commonly used metric is *F-measure* that combines the *Precision* and *Recall*.

$$F_\beta = \frac{(1+\beta^2)\cdot \text{Precision} \cdot \text{Recall}}{\beta^2 \text{Precision} + \text{Recall}}$$

F_1 measure gives equal weight to Precision and Recall.

3.3.3 *Results for Named Entity Categorization*

Table 3 shows the results for the Named Entity Categorization process using the dataset given in Section 4.3.1. We test our approach based on three set of features extracted from the Wikipedia articles:

- Features extracted from Heading Text (HT),
- Features extracted from Infobox Parameter Name (IP), and
- Features from both Heading Text and Infobox Parameter Name.

Table 3. Results of Named Entity Categorization

Evaluation Measure / Features	RI	Precision	Recall	F1
HT	0.943	0.746	0.746	0.746
IP	0.978	0.855	0.970	0.909
HT + IP	0.917	0.580	0.970	0.726

We notice that the features extracted from Infobox Parameter Name alone are sufficient to group Wikipedia categories that belong to a same concept. Almost 98% of category pairs are correctly categorized. The majority of Wikipedia articles in our dataset contain Inforbox information. In Wikipedia, most of the articles about person, company, location, movie and song contain Infobox. However, articles related to other concepts such as sports, dance, or martial arts are still lack of this structured information. In the case where no Infobox information is extracted, header text can be used as alternative features set in Named Entity Categorization problem.

From the results, we also found that our approach is not able to differentiate Wikipedia categories that contain many overlapping entities. For example, there are a number of entities appear in both categories *American_film_actors* and *American_film_directors* (many actors are also director). In our approach, we use Boolean vector (Section 4.1.1.2) to calculate similarity between two Wikipedia categories. The importance or weight of the features is not taken into consideration. Therefore, we get a high score for the above category pair and the two Wikipedia categories are categorized together.

3.3.4 *Results for Category Relation Learning*

Based on the Category Relation Learning approach described in Section 4.1.2, we calculate the relation for all concept pairs defined in Section 4.3.1. Table 4 below shows the weight for each concept pair (C_i, C_j).

From the results, it is obvious that concept pairs that are naturally closely related have higher score than others. For example, the weight for relation $R_{c3,c1}$ (0.8412) is higher than $R_{c3,c2}$ (0.3594), which mean that the concept '*Film*' is more related to the concept '*Actor*' than the concept '*Singer*'.

Table 4. Results of Category Relation Learning

Relation, $R_{ci,cj}$	Weight	Relation, $R_{ci,cj}$	Weight
$R_{c1,c2}$	0.2209	$R_{c2,c1}$	0.4078
$R_{c1,c3}$	0.6931	**$R_{c3,c1}$**	**0.8412**
$R_{c1,c4}$	0.0383	**$R_{c4,c1}$**	**0.2248**
$R_{c1,c5}$	0.4232	$R_{c5,c1}$	0.4194
$R_{c1,c6}$	0.2402	$R_{c6,c1}$	0.2971
$R_{c1,c7}$	0.0883	$R_{c7,c1}$	0.3417
$R_{c1,c8}$	0	$R_{c8,c1}$	0.1714
$R_{c2,c3}$	0.2929	**$R_{c3,c2}$**	**0.3594**
$R_{c2,c4}$	0.4389	**$R_{c4,c2}$**	**0.7422**
$R_{c2,c5}$	0.0890	$R_{c5,c2}$	0.1583
$R_{c2,c6}$	0.0509	$R_{c6,c2}$	0.1221
$R_{c2,c7}$	0.4354	$R_{c7,c2}$	0.7287
$R_{c2,c8}$	0.5243	$R_{c8,c2}$	0.5743
$R_{c3,c4}$	0.0112	$R_{c4,c3}$	0.0640
$R_{c3,c5}$	**0.6412**	$R_{c5,c3}$	0.7415
$R_{c3,c6}$	0.5503	$R_{c6,c3}$	0.6729
$R_{c3,c7}$	0.1350	$R_{c7,c3}$	0.0854
$R_{c3,c8}$	0	$R_{c8,c3}$	0.0855
$R_{c4,c5}$	**0.0213**	$R_{c5,c4}$	0.0317
$R_{c4,c6}$	0.0427	$R_{c6,c4}$	0.0488
$R_{c4,c7}$	0.2807	$R_{c7,c4}$	0.6878
$R_{c4,c8}$	0.2459	$R_{c8,c4}$	0.3164
$R_{c5,c6}$	0.2254	$R_{c6,c5}$	0.4183
$R_{c5,c7}$	0	$R_{c7,c5}$	0.0426
$R_{c5,c8}$	0.0631	$R_{c8,c5}$	0.0853
$R_{c6,c7}$	0.0244	$R_{c7,c6}$	0.0639
$R_{c6,c8}$	0.0489	$R_{c8,c6}$	0.1138
$R_{c7,c8}$	0.5983	$R_{c8,c7}$	0.2304

C1='Actor', C2='Singer', C3='Film', C4='Album', C5='Film_Director', C6='Film_Producer', C7='Song', C8='Music_Group'

3.3.5 *Results for Descriptive Features Selection*

We evaluate our Descriptive Feature Selection approach described in Section 4.2.1 by manually examining the descriptive features extracted by our algorithm. For each descriptive feature, we manually decide whether the feature is indeed descriptive of the articles in the Wikipedia category. Here, we test on five Wikipedia categories and calculate the Precision measure for each category. The result was shown in Table 5 below:

Table 5. Results of Descriptive Feature Selection

Wikipedia Categories	Precision
American_male_singers	0.5235
English_male_singers	0.6377
Scottish_male_singers	0.7414
Actors_from_New_York	0.6333
Actors_from_New_Jersey	0.2902

Table 6 below shows some of the highest rank features extracted by our approach. From the results, we found that our approach is able to extract more fine-grained attributes to describe the entities in a Wikipedia category. For example, the features '*New_York*', '*Brooklyn*' and '*Manhattan*' describe the place of birth of the entities in Wikipedia category '*Actors_from_New_York*' and are considered as important information to be annotated in a Wikipedia article.

Table 6. The Highest Rank Descriptive Features

Category	Desciptive Features
American_male_singers	United_States, Singer, Guitar, Musician, Rock_music, Songwriter, California
English_male_singers	England, Singer, Guitar, Musician, London, Pop_music, Rock_music, Songwriter, United_Kingdom
Scottish_male_singers	Scotland, Guitar, Rock_music, Glasgow, Musician, Folk_music, Vocals
Actors_from_New_York	United_States, Actor, Film, New_York, California, Television, Brooklyn, Los_Angeles, Manhattan
Actors_from_New_Jersey	United_States, New_Jersey, Actor, California, Newark, New_Jersey

3.4 Annotating Wikipedia Articles

In this section, we show how to annotate Wikipedia articles with the semantic information extracted using our approach. First, a Wikipedia article has to be converted into an XML document based on its logical structure [2][6]. Figure 4 shows an XML document for the Wikipedia page *Steven_seagal* (http://en.wikipedia.org/wiki/Steven_Seagal) from the INEX Wikipedia corpus [2].

For each Wikipedia article, we annotate the XML document with the conceptual information that describes the entity in the article. Based on our dataset (Figure 1), the page *Steven_seagal* is associated to three Wikipedia categories: *American_film_actors*, *American_film_directors*, *American_male_singers*. Therefore, the entity Steven Seagal is considered as belonging to the concepts '*Actor*', '*Film_Director*' and '*Singer*' (Table 1), and three conceptual tags <actor>, <film_director> and <singer> are added to the XML document. As the conceptual tags describe the classes or categories of the entity, they are added after the root tag <article> in the XML document (Figure 5). In this paper, the tag names are defined manually.

```
<?xml version = '1.0' encoding = 'UTF-8'?>
<article xmlns:xlink="http://www.w3.org/1999/xlink">
    <header>
        <title>Steven Seagal</title>
    </header>
    <body>
        <p><b>Steven Seagal</b> (born April 10, 1952) is
an American action movie actor, producer, writer, director...
        ...
        </p>
        <section>
            <title>Early years</title>
            <section>
                <title>Youth</title>
                <p>Segal was born in Lansing, Michigan,
where he lived until he was five years old.
                </p>
            </section>
        ...
</article>
```

Figure 4: XML Documents for page *Steven_seagal*

```
<?xml version = '1.0' encoding = 'UTF-8'?>
<article xmlns:xlink="http://www.w3.org/1999/xlink">
  <actor>
    <film_director>
      <singer>
        <header>
          <title>Steven Seagal</title>
        </header>
        <body>
          <p><b>Steven Seagal</b> (born April 10, 1952)
is an <born>American</born> action movie
<occupation>actor</occupation>,
<occupation>producer</occupation>, writer,
<occupation>director</occupation>, ...
          ...
          </p>
          <section>
            <title>Early years</title>
            <section>
              <title>Youth</title>
              <p>Segal was born in <born>Lansing,
Michigan</born>, where he lived until he was five years old.
              </p>
            </section>
          ...
      </singer>
    </film_director>
  </actor>
</article>
```

Figure 5: XML Documents for page *Steven_seagal* Annotated with Semantic Tags

For each entity, besides annotating the page that describes the entity, we also annotate other pages that link to the entity. For example, in both pages *Songs_from_the_Crystal_Cave* and *Under_Seige*, there are outgoing links to the page *Steven_seagal*. In this paper, we also intend to annotate the entity Steven Seagal in *Songs_from_the_Crystal_Cave* and *Under_Seige*. In order to identify the appropriate concept for the entity Steven Seagal in both pages, we take into consideration the relation weight of the concept pairs. The page *Songs_from_the_Crystal_Cave* is link to Wikipedia category *Steven_Seagal_albums*, which is belonging to the concept '*Album*'. Therefore, we calculate the relation weights for the concept pairs (*Album, Actor*), (*Album, Film_Director*) and (*Album, Singer*). From Table 3, the relation weights $R_{Album,Actor}$, $R_{Album,\ Film\_Director}$, and $R_{Album,\ Singer}$ are 0.2248, 0.0213, and 0.7422 respectively. This means that the concept '*Singer*' is more closely related to the concept '*Album*' compare to the other two concepts. Thus, the entity Steven Seagal is annotated with the concept '*Singer*' in the page *Songs_from_the_Crystal_Cave* (Figure 6). Similarly, we also identify the appropriate concept for Steven Seagal in the page *Under_Seige*. From out dataset, the page *Under_Seige* is link to Wikipedia category *American_films* and thus is considered as belonging to the concept '*Film*'. Here, we calculate the relation weight for $R_{Film,\ Actor}$(0.8412) $R_{Film,\ Film\_Director}$(0.6412), and $R_{Film,\ Singer}$ (0.3594). In this case, the concept 'Film' and 'Actor' are more closely related. Therefore, the entity Steven Seagal in the page *Under_Seige* is annotated with the semantic tag <actor> (Figure 7).

```
<?xml version = '1.0' encoding = 'UTF-8'?>
<article xmlns:xlink="http://www.w3.org/1999/xlink">
  <album>
    <header>
        <title>Songs from the Crystal Cave</title>
    </header>
    <body>
        <p><b> Songs from the Crystal Cave </b> is a 2005
album by <singer>Steven Seagal</singer>, his first.
        ...
        </p>
        ...
    </body>
  </album>
</article>
```

Figure 6: XML Documents for page *Songs_from_the_Crystal_Cave* Annotated with Semantic Tags

```
<?xml version = '1.0' encoding = 'UTF-8'?>
<article xmlns:xlink="http://www.w3.org/1999/xlink">
  <film>
    <header>
        <title>Under Seige</title>
    </header>
    <body>
        <p><b>Under Seige </b> is a 1992 action film.
Directed by <film_director>Andrew Davis</film_director>,
it stars <actor>Steven Seagal</actor> as a formal Navy Seal
        ...
        </p>
        ...
    </body>
  </film>
```

Figure 7: XML Documents for page *Under_Seige* Annotated with Semantic Tags

In addition to conceptual information, we also annotate the XML document with descriptive features of the entity. For each descriptive feature, we examine if the feature occurs as Wikipedia Infobox parameter value. As the Infobox parameter name indicates the semantics of the attribute, we use the parameter name as our tag name. For example, the feature '*United_States*' is the value for the parameter name '*born*', whereas the feature '*Actor*' is associated with the parameter name '*occupation*'. In the page *Steven_seagal*, the features 'America' and 'Lansing, Michigan' are annotated with the semantic tag <born>, whereas 'actor', 'producer' and 'director' are annotated with <occupation> (Figure 5).

4. APPLICATIONS TO STRUCTURED RETRIEVAL

It has been shown that annotating XML documents with semantic information such as person, organization, location and etc will improve the precision of search results [4][6][11]. We have not yet done a thoroughly evaluation of the retrieval performance using our annotated documents. However, as our approach aims to explicitly annotate the conceptual information and descriptive information of an entity in the Wikipedia article, it is likely that these annotations can help to enhance the retrieval results.

For example, given the query <singer>Steven Seagal</singer> (Section 3.2), without the semantic annotations in the XML documents, all documents related to Steven Seagal, including both *Songs_from_the_Crystal_Cave* and *Under_Seige*, will be returned to the user. By adding the semantic tag <actor> and <singer> to the documents, only documents related to Steven Seagal's music or album (*Songs_from_the_Crystal_Cave*) will be returned. This give more focused results to the users and enable them to locate the information they are looking for more quickly.

5. RELATED WORK

There have been several attempts to annotate documents with semantic tags. Zwol [11] describes a named entity detection approach based on a combination of heuristically derived rules and pre-defined thesauri. Four types of named entities, i.e. person, company, organization and location, are detected. The derived set of named entities is used to annotate the Reuters collection.

Schenkel [6] presents a system, YAWN (Yet Another Wikipedia Annotation) that converts the Wikipedia collection into an XML corpus with semantically rich and self-explaining tags. YAWN annotates the entities in Wikipedia articles with concepts from WordNet. Three types of information from Wikipedia corpus are used in the system: categorical information, lists of similar pages, and Infobox.

Our approach is different from other works in the sense that we aim to identify the semantics or concepts of an entity in different context. Instead of annotating an entity with general class or concept such as person, location, organization etc, we intend to annotate the entity with more fine-grained concepts. For example, in Figure 5, the entity '*Lansing, Michigan*' represents the place of birth of Steven Seagal, thus it is annotated with the semantic tag <born> rather than with its general concepts such as <city> or <state>.

6. CONCLUSION

In this paper, we present an approach to automatically identify the semantic information in Wikipedia articles and explicitly annotate the information with semantic tags. We propose an unsupervised learning approach to learn the concept of the entities in Wikipedia articles and also extract the entities' descriptive features from the articles.

We test our approach on a small-scale of dataset and found that the results are favorable. For future work, we plan to extensively evaluate the quality of the semantic information extracted using the proposed approach and also examine the effectiveness of the semantic structure in improving the retrieval performance, especially the precision value.

7. REFERENCES

[1] Bray, T. Paoli, J., Sperberg-McQueen, C.M. and Maler, E. The Extensible Markup Language (XML) 1.0 (Fifth Edition) W3C Recommendation. Nov 2008.

[2] Denoyer, L. and Gallinari, P. 2006. The Wikipedia XML Corpus. ACM SIGIR Forum, 40, 1, (June 2006). 64-69

[3] Fuhr, N. Lalmas, M., Malik, S. and Kazai, G. Advances in XML Information Retrieval and Evaluation. In Proceedings of the Fourth Workshop of the Initiative for the Evaluation of XML Retrieval (Dagstuhl, Nov 2005). INEX'05.

[4] Graupmann, J., Schenkel, R. and Weikum, G. The Spheresearch Engine for Unified Ranked Retrieval of Heterogeneous XML and Web Documents. In Proceedings of the 31th International Conference on Very Large Data Bases (Trondheim, Norway, 2005). VLDB'05. 529-540.

[5] Salton,G., Wong, A., and Tang, C.S. 1975. A Vector Space Model for Information Retrieval. Communications of the ACM. 18, 11 (Nov. 1975), 613-620.

[6] Schenkel, R., Suchanek, F. and Kasneci, G. 2007. YAWN: A Semantically Annotated Wikipedia XML Corpus. In Proceedings of Datenbanksysteme in Business, Technologie und Web. BTW'2007. 277-291.

[7] Schlieder, T., Meuss, H. Querying and Ranking XML Documents. American Society for Information Science and Technology. 53, 6 (2002), 489-503.

[8] Trotman, A. Searching Structured Documents. Information Processing and Management. 40, 4 (2004), 619-632.

[9] Yu, J., Thom, J.A., and Tam, A. 2007. Ontology Evaluation using Wikipedia Categories for Browsing. In Proceedings of 16th ACM Conference on Information and Knowledge Management (Lisbao, Portugal, 2007). CIKM'07. 223-232

[10] Vercoustre, A-M., Thom, J.A., Pehcevski, J. Entity Ranking in Wikipedia. In Proceedings of the 23 ACM Symposium on Applied Computing (Fortaleza, Veara, Brazil, 2008). SAC'08. 1101-1106.

[11] van Zwol, R. and van Loosbroek, T. Effective Use of Semantic Structure in XML Retrieval. In Proceedings of the 29th European Conference on IR Research (Rome, Italy, 2007). ECIR'07. 621-628