Dynamic Classification for Search Results of Ontology-based Contents

Yongjun Zhu*, Xiuyan Cai*, Wooju Kim**
izhu@yonsei.ac.kr, sooyeonchae@yonsei.c.kr, wkim@yonsei.ac.kr

Abstract

Ontology is a framework for organizing information. It describes resources which are the objects of information systems and their relationships definitely and specifically. This characteristic of ontology enables intelligent information searching in many areas. Although the efficiency of searching itself has been improved, the reasonable methodology for classifying search result remains as an issue. Search result often consists of many items, so an efficient classification system can help users save the time reviewing items irrelevant to his purpose. Furthermore, most present classification systems are not dynamic; they are based on pre-defined categories. In this paper, we propose a novel classification methodology which can dynamically classify search result of ontology-based contents, specifically movies in considering of their similarities.

1. Introduction

Ontology's superior ability of describing resources as well as their relationships and the development of semantic web technology and ontology description languages facilitate many information systems adopting ontology as the structural framework for organizing information due to the fact that intelligent information searching is available with such technologies. Take the movie searching system as an example, if movie information is stored in ontology, the system can process complicated queries such as "action movies of people who starred in Mission: Impossible" and shows

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2011-0027007).

This work (research) is financially supported by the Ministry of Knowledge Economy (MKE) and Korea Institute for Advancement in Technology (KIAT) through the Workforce Development Program in Strategic Technology

* Yonsei University Department of Information & Industrial Engineering Master's course

** Yonsei University Department of Information & Industrial Engineering Professor

people the search result. In many systems, the search result is often classified based on pre-specified classification rules. Many movie sites like IMDB [1] classifies movies based on fixed condition such as genre, country even though the search result is always different. In the aforementioned example we cannot use genre as a classification criteria because all the movies in search result have the same genre-action. In this respect, classification criteria must be changed based on search result, otherwise the classification itself is meaningless.

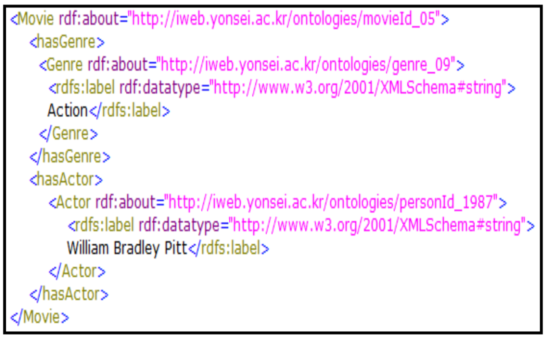
Be faced with these issues, we propose our novel methodology which can dynamically classify ontology-based search result. The object of our system is movie. Of course, object can be any contents as long as the contents are well defined in ontology. We compare all movie properties (including genre, country, director, producer etc.) and choose best three properties and classify based on the three properties. Take into consideration that not all users have the same perspective, we choose three properties and users can choose one which they prefer.

The organization of this paper is as follows. Section 2 lists some knowledge terms used in this paper such as OWL 2 and information gain. Section 3 presents our detailed methodology. Section 4 demonstrates some examples and evaluates our methodology. We give the conclusion and future work in the last section.

2. Related Work

2.1 OWL 2

OWL2 Web Ontology Language (OWL 2) is an ontology language for the semantic web. OWL 2 ontologies provide classes, properties, individuals, and data values and are stored as semantic web documents [2]. The movie information used by our system is stored in ontology using OWL 2.



```
<Movie rdf:about="http://iweb.yonsei.ac.kr/ontologies/movieId_05">
   <hasGenre>
     <Genre rdf:about="http://iweb.yonsei.ac.kr/ontologies/genre_09">
       <rdfs:label rdf:datatype="http://www.w3.org/2001/XMLSchema#string">
       Action</rdfs:label>
     </Genre>
   </hasGenre>
   <hasActor>
       <Actor rdf:about="http://iweb.yonsei.ac.kr/ontologies/personId_1987">
           <rdfs:label rdf:datatype="http://www.w3.org/2001/XMLSchema#string">
           William Bradley Pitt</rdfs:label>
       </Actor>
   </hasActor>
</Movie>
```

<Figure 1> OWL 2 syntax

<Figure 1> shows OWL 2 syntax. It means there is a movie (movieId_05), its genre is Action (genre_09) and it has an actor called William Bradley Pitt (personId_1987). Our information about movie is described in the same way as above format in ontology.

2.2 Information Gain

In information theory, entropy is a measure of the average information content one is missing when one does not know the value of the random variable [3].

It is a measure of unpredictability, high entropy means the outcome is unpredictable and zero entropy means you always know the outcome. In his paper, Shannon [3] denoted the entropy H of a discrete random variable X with possible values $\{x_1, \cdots, x_n\}$ as

$$\mathrm{H}(X) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i) \qquad (1)$$

where $\mathrm{p}(x_i)$ is the probability mass function of outcomex_i. There is another term called conditional entropy. Conditional entropy of two events X and Y taking values x_i and y_i respectively is defined as

$$\mathrm{H}(X|Y) = \sum_{i,j} p(x_i, y_j) \log \frac{p(y_j)}{p(x_i, y_j)} \qquad (2)$$

where $p(x_i, y_j)$ is the probability that $\mathrm{X} = x_i$ and $\mathrm{Y} = y_j$. This quantity is the amount of randomness in the random variable X given that you know the value of Y.

Information gain is a non-symmetric measure of the difference between two probability distributions [4]. The expected information gain is the change in information entropy from a prior state to a state that takes some information. Information gain $\mathrm{IG}(X, Y)$ is defined as

$$\mathrm{IG(X,Y)} = \mathrm{H}(X) - H(X|Y) \qquad (3)$$

This means the reduction in the entropy of X achieved by learning the state of the random variable Y.

3. Dynamic classification using information gain

3.1 Properties of movie

Movie has some properties such as actors, producers, etc. We considered 11 properties of movie; they are listed in table 1.

<Table 1> Properties of movie

No.	Properties	Examples
1	genre	Action
2	country	Korea
3	director	personId0001
4	starring	personId0002
5	producer	personId0003
6	writer	personId0004
7	cinematographer	personId0005
8	editor	personId0006
9	artDirector	personId0007
10	musicDirector	personId0008
11	otehrCrew	personId0009

Each property can have more than one property value. Some date type properties such as “initial release date” are not considered in this paper. The object of this paper is to choose three properties that best represent the characteristic of search result among these 11 properties.

3.2 Calculation of Information Gain

To achieve our goal which is finding best three properties, we use information gain as the criteria of choosing these properties. As we mentioned in section 2, information gain means the reduction in the entropy of property A achieved by learning the information about property B. So, in other words, our goal is finding properties which maximize information gain. Properties with high information gain mean that they represent the search result well, so they can be good classifiers. Calculating information gain of each property can be divided into 4 steps.

<Table 2> Steps of calculating information gain

No.	Description
Step 1	Calculate entropy of each property
Step 2	Target one property and calculate information gain of other properties.
Step 3	Normalize information gain using information gain ratio and figure out the sum of information gain ratio by targeting different properties.
Step 4	Rank the sum of information gain ratio of each property and choose 3 properties which have bigger information gain ratio than others.

Step 1:

We calculate entropy H of property p_x as follows:

$$\mathrm{H}(p_x) = -\sum_{i=1}^{n} \frac{num(p_x^v(i))}{num(I)} \log_2 \frac{num(p_x^v(i))}{num(I)} \quad (4)$$

where p_x^v denotes the value of property p_x, n denotes the number of distinct values of property p_x, $\mathrm{p}_x^v(i)$ denotes i th value of property p_x, I denotes total instances, and $\mathrm{num(I)}$ and $\mathrm{num}(\mathrm{p}_x^v(i))$ denote the number of instances I and the number of instances who have the $\mathrm{p}_x^v(i)$ as the value of property p_x. After calculating entropy of each property, we calculate information gain of property by targeting another property.

In order to obtain information gain conditional entropy is need. Conditional entropy $\mathrm{H}(p_x|p_y)$ can be calculated as:

$$\mathrm{H}(p_x|p_y) = -\sum_{i=1}^{n}\sum_{j=1}^{m} \frac{num(\mathrm{p}_y^v(i)\,)}{num(I)} \frac{num(\mathrm{p}_x^v(j))}{num(\mathrm{p}_y^v(i))} \log_2 \frac{num(\mathrm{p}_x^v(j))}{num(\mathrm{p}_y^v(i))}$$

(5)

where n denotes the number of distinct values of property p_y, m denotes the number of distinct values of property p_x when divided by values of property p_y.

Step 2:

After the calculation of $\mathrm{H}(p_x)$ and $\mathrm{H}(p_x|p_y)$ information gain $\mathrm{IG}(p_x, p_y)$ can be calculated as follows:

$$\mathrm{IG}(p_x, p_y) = \mathrm{H}(p_x) - \mathrm{H}(p_x|p_y) \quad (6)$$

Step 3:

The value of information gain depends on the number of distinct values of property, so we should normalize the value. To obtain this we use the concept of information gain ratio [5]. We calculate split information as follows:

$$\mathrm{SI}(p_x, p_y) = -\sum_{i=1}^{n} \frac{num(\mathrm{p}_y^v(i))}{num(I)} \log_2 \frac{num(\mathrm{p}_y^v(i))}{num(I)} \quad (7)$$

Then $\mathrm{GR}(p_x, p_y)$ can be obtained as follows:

$$\mathrm{GR}(p_x, p_y) = \frac{\mathrm{IG}(p_x, p_y)}{\mathrm{SI}(p_x, p_y)} \qquad (8)$$

After calculating information gain ratio of all properties by targeting different properties, we sum up the total information gain ratio of each property.

$$\mathrm{GR}(\mathrm{P}, p_x) = \sum_{i=1}^{n-1} GR(p_i, p_x) \qquad (9)$$

where n is the number of properties.

Step 4:

The last step is to rank properties by obtained information gain ratio and choose top 3 properties and classify. Instances that have same property values are classified together.

4. Experiment

In order to verify our methodology, we did some experiments using the search results of below three example queries. We developed search system used in this experiment before writing this paper using Java programming language. The data was stored in ontology and Jena was used when dealing with ontology. A* algorithm was also used when searching.

<Table 3> Example queries

No.	Queries
1	dogani starring actor starring movie
2	action movie
3	RyuSeungwan movie

Search result of the first query “dogani starring actor starring movie” includes movies of people who also starred in movie “dogani”, query “action movie” searches for movies whose genre are action. “RyuSeungwan” is the name of Korean director and the search result of query “RyuSeungwan movie” includes movies which are directed by director “RyuSeungwan”. The scores of information gain ration of properties of each query are in Table 4, Table 5 and Table 6.

<Table 4> Scores of “dogani starring actor starring movie”

Order	Properties	Gain Ratio
1	director	7.709653232
2	music	7.699818535
3	editor	7.366900388
4	cinematographer	6.943126149
5	writer	6.12311236
6	artDirector	6.1065219
7	genre	5.807588506
8	producer	5.023904812
9	starring	3.11056763
10	country	0
11	otherCrew	-

Property “otherCrew” has no value because the website we used when collecting data do not offer information about “otherCrew” of some instances. So, in this query “director” is the best property which classifies the search result best.

<Table 5> Scores of "action movie"

Order	Properties	Gain Ratio
1	director	8.174155171
2	editor	7.491425148
3	cinematographer	7.436420332
4	music	7.332241494
5	country	7.262297637
6	otherCrew	6.090414058
7	artDirector	6.02607833
8	writer	5.68784101
9	producer	5.618449606
10	genre	5.206631528
11	starring	3.456767926

In this case, "director" is the best property again.

<Table 6> Scores of "RyuSeungwan movie"

Order	Properties	Gain Ratio
1	cinematographer	8.174155171
2	artDirector	7.491425148
3	editor	7.436420332
4	music	7.332241494
5	director	7.262297637
6	writer	6.090414058
7	genre	6.02607833
8	otherCrew	5.68784101
9	producer	5.618449606
10	starring	5.206631528
11	country	3.456767926

In this query, property "cinematographer" is the best property, so search result will be classified based on this property.

5. Conclusion

In this paper, we proposed dynamic classification methodology for search result of ontology-based contents. Though we focused our paper on ontology-based contents, it can be used by any system. We used information gain as a key concept to choose best property and the results of experiment are quite satisfactory. The number of classification is neither too big nor too small and one may have noticed that properties which should not have high score such as "starring" in the first query, "genre" in the second query and "director" in the third query have low scores. For example, search result of the second query are movies whose genre are action, so property "genre" should have low score so that the search result can be classified based on other properties. The property "genre" in Table 5 was ranked in 10th order. In this sense we can conclude that our methodology is reasonable.

References

[1] http://www.imdb.com/

[2] http://www.w3.org/TR/owl2-profiles/

[3] C.E. Shannon, "A Mathematical Theory of Communication", Bell System Technical Journal, Vol. 27, pp. 379-423, 623-656, 1948

[4] Kullback, S., Leibler, R.A., "On Information and Sufficiency", Annals

of Mathematical Statistics, Vol.22, No. 1, pp.79-86, 1951.

[5] Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Introduction to Data Mining, Addison Wesley, Boston, Massachusetts, 2005.