Broadcasting Program Search based on Ontology Search and Ranking Technique

Jungmin Kim[1], Gukbo Kim[1] and Hyunsook Chung[2]
[1]*Dept. of Computer Engineering, Daejin University*
{jmkim, kgb}@daejin.ac.kr
[2]*Dept. of Computer Engineering, Chosun University*
hsch@chosun.ac.kr (Corresponding author)

Abstract

*Now, user can watch program contents selectively through IPTV service at anytime. User, however, has the difficulty in searching the desired program contents because of inefficient retrieval interface and content annotation of IPTV. Our search and ranking method suggests broadcasting programs relevant to the queried broadcasting program based on their semantic knowledge structures, which describe core concepts and relationships extracted from the contents of broadcasting programs. Program search technique is performed based on the revised tf*idf. We define four ranking measurements, topic label coverage, topic completeness, relation complexity, and topic density, to rank the retrieved programs according to their ranking scores.*

Keywords: *IPTV, ontology, broadcasting program, searching, ranking*

1. Introduction

Internet Protocol Television called IPTV is defined as internet television service such as television, video, audio, text, graphics, and data delivered over IP based networks [1]. IPTV enables users to select live broadcasting channels, time-shifted programs, and video on demand. However, IPTV provides so many channels, programs, and multimedia data to users who have difficulty in choice and retrieve desired programs. For example, IPTV service providers in Korea, such as Skylife, KT Qook, and SK Broad&, deliver electronic program guide services(EPG) stored in their set-top boxes. However, current channel or program search interfaces of EPGs have problems in efficiency, correctness and semantic relation between program contents. Users should move program categories, like drama, sports, documentary, etc, iteratively. In addition, a few button presses in a remote control is required to reach a certain broadcasting program.

In this paper, we tackle this searching problem of IPTV through design of the broadcasting program ontology and search and ranking method. Our broadcasting program ontology conceptualizes interesting scenes semantically and defines meaningful relations between them. This ontology consists of schema ontology, metadata ontology and domain ontology in the layered structure. Ontology search problem should be solved to support semantic-based broadcasting program search. Ontology search techniques, i.e. Swoogle [2], Ontokhoj [3], and OntoSearch [4], similar to traditional web search engines have been developed and proposed to support the reuse of existing ontologies. They search ontologies by keyword or HTML document from ontology libraries. In section 2, we introduce more detail function of above ontology search techniques.

Keyword-based or document-based search is not sufficient to detect the best appropriate ontology for matching with the given ontology because only several keywords are not appropriate to be the representative of semantic information described in ontologies. This paper presents a new system framework incorporating search, matching and merging processes for broadcasting program domain ontologies.

2. Related Work

Previous researches related to our study have been studied about integration and retrieval of multimedia data efficiently. Song, et. al., [5] propose an mapping ontology of MPEG-7 and TV-Anytime metadata, which defines the correspondence relations between entities of two metadata.

Ardnt, et. al., [6] and Issac, et. al., [7] construct generic ontologies extending MPEG-7 metadata through definition of new elements, which can be used in various multimedia application services. Tsinariki, et. al., [8] and Bellekens, et. al., [9] transform MPEG-7 metadata into OWL ontology to support indexing and searching of multimedia files. They construct base ontologies based on MPEG-7 MDS semantic part and build domain ontologies over the base ontologies.

However, these previous researches have focus in MPEG-7 and multimedia data like video and audio files only. They include a small part of TV-Anytime metadata, i.e. genre classification and program description, rather than whole part. Thus, these approaches are inadequate to index and search broadcasting programs on IPTV because they do not consider features of broadcasting program contents.

There are a few studies related to the development of ontology search engines in order to support knowledge reuse. Swoogle [2] is a crawler-based indexing and retrieving system for RDF or OWL documents. OntoKhoj [3] is a Semantic Web portal designed to simplify ontology engineering process. It is based on algorithms used for searching, aggregating, ranking and classifying ontologies in Semantic Web. OntoKhoj crawler fetches the RDF documents according to the physical links and then aggregate several RDF segments belong to same logical URI but physically present at different locations into a single ontology.

OntoSearch [4] is another ontology search engine using Google APIs and hierarchy visualization technique. It allows users to search certain types of ontology files by keywords only. Swoogle and OntoSearch are initiative ontology search engines but they have weakness due to keyword-based querying. OntoSelect [10] is a web-based ontology library that collects, analyzes, and orgarnizes ontologies and allows searching as well as browsing of ontologies according to size, representation format, connectedness and human languages used for class labels. OntoSelect provides ontology search based on one or more keywords and a HTML document.

We have a different perspective from previous approaches because our ontology search and ranking modules are tightly coupled to matching process to find the best candidate. We takes a whole domain ontology as query input in order to find other relevant broadcasting program domain ontologies, which have higher similarity in terms of syntactic and semantic structure, from the collection of domain ontologies.

3. Ontology Structure

The structure of broadcasting ontology is three-layered architecture which is composed of TV-Anytime ontology, domain ontologies, and top-level ontologies. TV-Anytime ontology, which is the metadata ontology, conceptualizes description of

broadcasting programs and VOD multimedia files and is located in the bottom layer. Domain ontology is a set of independent ontologies, which conceptualize contents of broadcasting programs according to their genres like drama, news, sports, documentary, etc.

Domain ontology is located in middle layer and has links to top-level ontology and TV-Anytime ontology. Top-level ontology also is a set of common ontologies like ABC ontology, Time ontology, and Geography ontology. We define a mapping table that includes correspondence relations between elements of XML schema type and OWL constructs to convert TV-Anytime metadata into our metadata ontology. For example, *ComplexType* of XSD(XML Schema Datatype), is used to describe complex entities, such as category, program, producer, channel, etc. Thus, *ComplexType* is transformed into *Class* of OWL. In addition, *SimpleType* of XSD can be mapped to *DatatypeProperty* of OWL.

Domain ontology, i.e. *Drama ontology*, *Sports ontology*, and *Documentary ontology*, conceptualizes the contents of broadcasting programs through core concept identification, term definition, and semantic network creation. Domain ontology enables users to search certain broadcasting programs or related contents to watching program. In this section, we present domain ontology construction methodology. Broadcasting programs delivered from IPTV have web pages, which describe synopsis, character, credit, and so on. This basic information may be provided to users using TV-Anytime metadata. We use the auxiliary information of programs in order to extract knowledge of contents of programs.

Our domain ontology construction process is composed of two main phases, core concept identification and semantic network construction phase. Figure 1 shows process of core concepts extraction from web pages of broadcasting programs.

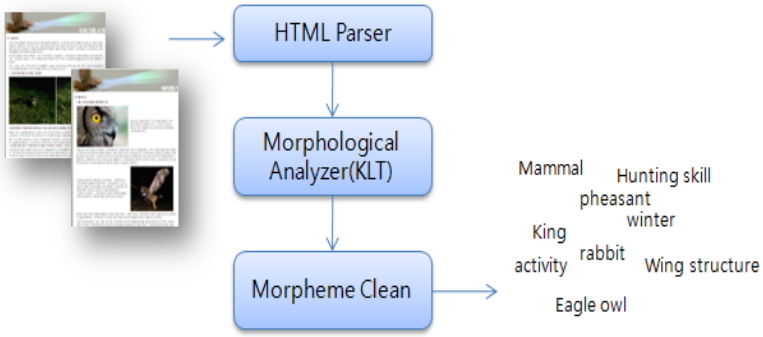

Figure 1. Extraction of Core Concepts from Web Pages

HTML parser parsing collected web pages related to broadcasting programs in order to remove unnecessary data like images, symbols, and numbers, and extract textual data. Textual data extracted from several web pages are merged in a text file and passed into morphological analyzer to identify actual morphemes. Domain experts examine actual morphemes manually in order to identify core concepts of programs. This work is processed in morpheme clean step. Next phase create the semantic network of core concepts and instances of the concepts. Semantic relations among concepts include *superclass-subclass* relations and domain specific relations, i.e. *isMammal*, *isKindOf*, *liveIn*, etc.

4. Search and Ranking

In this section, we describe our ontology search and ranking method, which is used to find relevant broadcasting program. The system architecture is represented in Figure 2.

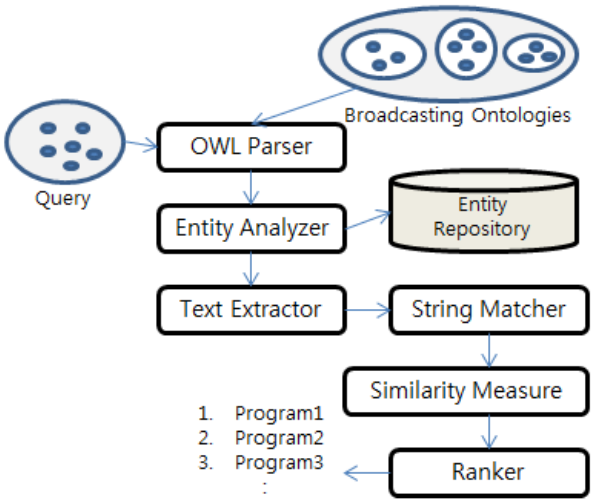

Figure 2. The System Architecture for Searching and Ranking Broadcasting Ontologies

Our ontology search is the process of retrieving a list of ontology potentially relevant to a given ontology. We use a single ontology instead of a few keywords as query data to achieve the processing of ontology matching and merging successfully. In addition, we develop a ranking method considering concept structure and semantic relations to find the best appropriate candidate. Because keyword-based query may be insufficient to determine the correct sense of terms, ontology search employs auxiliary knowledge resources, like WordNet or Open Directory, and traverses ontology graphs [3].

This complex query processing requires lots of time to retrieve ontology files. For efficiency of ontology search, we use only exact and partial string matching between terms extracted from topic labels. We check semantic structure of terms in the ranking and matching processes.

In this paper, we use the normalized frequency-based vector model instead of classic TF*IDF based vector model because document frequency, which means the number of documents containing term *x*, is not appropriate to ontology search. In other words, document frequency is used to assign a low weight to too common terms in IR systems. But the number of ontology containing common terms is not important factor in the ontology search process. The normalized frequency of term *x*, which is a weight of term *x*, in an ontology is given by the following expression.

$$\textit{Term weight } w(x) = \left(\frac{tf(x)}{max\ tf(y)}\right) \quad (1)$$

From above expression, *max tf(y)* denotes the maximum frequency of term *y*. Thus the expression computes a term ratio for each term in ontology. We create a term vector based on the normalized frequencies of terms. Two term vector of two ontologies are compared to measure the similarity between them to assign the similarity value(score) to the pair of ontologies.

Our ranking module takes the list of ontologies and ranks them according to their scores produced by applying four different measures such as concept completeness, relation complexity and concept density. The ranking measures produce three scores for each of the retrieved ontologies. These scores must be weighted and combined to generate a single final ranking score for each of ontologies. Figure 3 represents our proposed ranking algorithm.

```
Algorithm. Ranked List Generation

Input. The list of retrieved domain ontologies
Output. The ranked list of domain ontologies

Procedure RankingPrograms(OntoList, SimVaueSet)
var
    RankList := NULL
    ScoreSet := NULL
begin
    For each ontology in OntoList
        cov := CompletenessMeasure()
        com := ComplexityMeasure()
        den := DensityMeasure()
        score := CompositeMeasure(SimValueSet(i), cov, com, den)
        InsertToScoreSet(score)
    Next
    RankOutput(OntoList, ScoreSet, RankList)
end
```

Figure 3. The Ranking Algorithm based on Three Measurements

The concept completeness measures the level of conceptualization of each matched concept. Generally, the level of conceptualization of a concept depends on the number of properties and relations with other concepts. This measurement assigns higher score to a concept having a relatively large number of properties. The score of concept completeness measurement is computed by the following expressions (2) and (4).

$$tc(i) = w*|ICi| + (1-w)*|ECi| \quad (3)$$

$$TC = \sum_{i=1}^{n} tc(i) \quad (4)$$

In expression (3), *tc(i)* denotes the concept completeness score of the concept *i*. *IC* and *EC* denote the number of internal properties and external properties respectively. The relation complexity measures how well concepts are interconnected based on sematic relations, such as superclass, subclass, association, and so on. This measurement, like the concept completeness measurement, also represents the level of conceptualization of ontology. We identify five types of relations. They are superclass, subclass, instanceOf, sibling and association, which exist between concepts in ontology. We compute the concept relation complexity of the matched concepts using the following two expressions.

$$trc(i) = w1*Super(i)+w2*Sub(i)+w3*I(i)+w4*A(i)+w5*sib(i) \quad (5)$$

$$TRC = \sum_{i=1}^{n} trc(i) \quad (6)$$

Lastly, the concept density measures how many intermediate concepts between matched concepts are existed. We define concept density measurement to find the best

matching candidate from the list of ontologies. The best matching candidate denotes an ontology, which can extend the semantic structure and enhance semantic quality of query ontology after matching and merging. Thus, matched concept pair having longer distance has higher concept density score, which means better conceptualization. Following expression (7), (8) and (9) are evaluated to produce the score.

Let $t_i \rightarrow t_j$ be a shortest path between concepts t_i and t_j. Distance between the concepts can be computed using the expression (7).

$$dist(t_i,t_j) = \begin{cases} length(t_i \rightarrow t_j), & if\ t_i \neq t_j \\ 1, & if\ t_i = t_j \end{cases} \quad (7)$$

If query ontology has two concepts t_x and t_y that they are matched t_i and t_j to each other, the concept density score of a pair of concepts t_i and t_j can be computed by evaluating following expression (8). In this expression, $dist_q(t_x, t_y)$ denotes distance between the concepts t_x and t_y in query ontology.

$$td(t_i,t_j) = \frac{dist(t_i,t_j)}{dist_q(t_x,t_y)} \quad (8)$$

Let $dist_k$ be the concept density score of a pair of two concepts in a retrieved ontology. Following expression (9) produce the accumulated score of each concept density score for an ontology.

$$TD = \sum_{k=1}^{m} dist_k \quad (9)$$

5. Experiment

To evaluate the performance of our approach, we prepared three groups of 41 broadcasting program contents which are documentary programs dealing with different subjects, such as nature, science, animal, and culture. Table 1 represents a partial list of the experiment data. We classified the collected documentary programs into three groups according to their relevance in contents. Group A has documentary programs which describe similar subjects and contain same a few keywords as given programs used as query data. Group B has documentary programs which represent different subjects but same a few keywords as query data. Group C has documentary programs, which contain different subjects and keywords to query data.

Table 1. A Partial List of Experiment Data

Group	No.	Program Title	Subject
A	1	Kingdom of Animal	Animal
	4	Wilde Beast of Africa	Animal
	8	Chimpanzees of Tanganyika	Animal
B	9	Wildebeest Migration	Animal
	15	Queen of Africa	Culture
C	17	Insight Asia: Noodle Road	Culture
	35	Into Science	Science

We classified the collected documentary programs into three groups according to their relevance in contents. Group A has documentary programs which describe similar subjects and contain same a few keywords as given programs used as query data. Group B has documentary programs which represent different subjects but same a few keywords as query data. Group C has documentary programs which contain different subjects and keywords to query data. We used an episode of kingdom of animal series and people of Kyrgyzstan as query data. We measured the performance of searching and ranking using precision, recall, and f-measure measurements.

Figure 4 represents experiment results of searching and ranking for two query data. We compared our approach to keyword-based approach which is used in previous related works and in current IPTV remote controllers. We extracted less than 5 keywords from main concepts of query data ontologies, such as lioness, hunting, live in troops, etc.

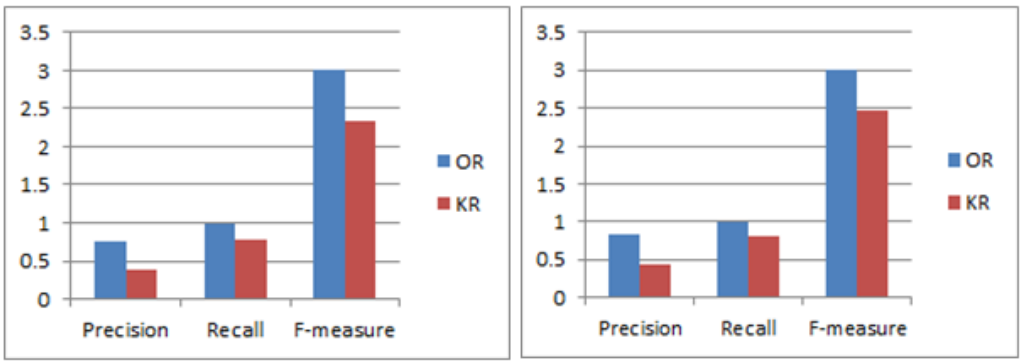

Figure 4. Comparison between our Approach (OR) and Keyword-based Approach (KR) for Kingdom of Animal Episode (a) and People of Kyrgyzstan (b)

From experimental result and performance evaluation, we found that our approach retrieved all relevant programs for given query data ontologies in spite of 80% precision rate. The cause of lower precision than recall is that some documentary programs irrelevant to query data are included in the search result because they have similar keywords to query data. Keyword-based approach has lower precision that our approach because irrelevant programs, which have same keywords but different subjects to query data, are included in the search result. This means that a few keywords only cannot represent core concepts of the contents of broadcasting programs.

6. Conclusion

In this paper, we present a new system for searching and ranking broadcasting programs based on broadcasting ontologies. Our searching method has the process of retrieving a list of broadcasting domain ontologies potentially relevant to a given query data ontology. Our experiments for searching documentary programs prove that ontology-based searching is more precise than keyword-based searching in comparison of contents semantically. To speed up ontology-based broadcasting program search, the precise statistical data of each of ontologies must be stored and used. Thus our future work is the development of automatic searching process and reuse of statistical data.

Acknowledgements

This work was supported by the Korea Research Foundation Grant funded by the Korean Government (No. 2011-0014613).

References

[1] ATIS IPTV Exploratory Group Report and Recommendation to the TOPS Council, http://www.atis.org/tops/IEG/ATIS_IPTV_EG_RPT_final.pdf.

[2] L. Ding, T. Finin, A. Joshi, R. Pan, R. S. Cost, Y. Peng, P. Reddivari, V. C. Doshi and J. Sachs, "Swoogle: A semantic web search and metadata engine", In Proceedings of the 13th ACM Conf. on Information and Knowledge Management, **(2003)**.

[3] C. Petel, K. Supekar, Y. Lee and E. K. Park, "OntoKhoj: A Semantic Web Portal for Ontology Searching, Ranking and Classification", In Proceedings of the Workshop On Web Information And Data Management, **(2003)**, pp. 58-61.

[4] Y. Zhang, W. Vasconcelos and D. Sleeman, "Ontosearch: An ontology search engine", In Proceedings of the Int. Conf. on Innovative Techniques and Applications of Artifial Intelligence, **(2004)**, Cambridge, UK..

[5] C. -H. Song and S. -J. Yoo, "MPEG-7 MDS and TV-Anytime-based Ontology for Semantic Retrieval of Multimedia Data", Journal of The Korean Society of Broadcasting Engineers, vol. 11, no. 1, **(2007)**, pp. 42-53.

[6] R. Ardnt, R. Troncy, S. Staab, L. Hardman and M. Vacura, "COMM: Designing a Well-Founded Multimedia Ontology for the Web", Lecture Notes in Computer Science, vol. 4825, **(2007)**, pp. 30-43.

[7] A. Isaac and R. Troncy, "Designing and Using an Audio-Visual Description Core Ontology", International Workshop on Core Ontologies in Ontology Engineering, **(2004)**.

[8] C. Tsinaraki, P. Polydoros, N. Moumoutzis and S. Christodoulakis, "Coupling OWL with MPEG-7 and TV-Anytime for Domain-Specific Multimedia Information Integration and Retrieval", In Proceedings of RIAO, **(2004)**.

[9] P. Bellekens, K. van der Sluijs, G.-J. Houben and L. Aroyo, "On-the-fly Data Integration for Personalized Television Recommender Systems", 8th International Conference on Web Engineering, **(2008)**.

[10] P. Buitelaar, T. Eigner and T. Declerck, "OntoSelect: A Dynamic Ontology Library with Support for Ontology Selection", In Proceedings of the International SemanticWeb Conference. Hiroshima, Japan, **(2004)**.

Authors

Jungmin Kim

Jungmin Kim is a professor in the Department of Computer Engineering at Daejin University, Pocheon, Korea. He received a B.S. degree and a M.S. degree in Computer Science in Hongik University in 1992 and 1994 respectively. He received a Ph.D. degree in Computer Engineering from Seoul National University in 2007. His research interests include Semantic Web, Semantic Information Processing, IPTV, and Smart Learning.

Gukbo Kim

Dr. Guk-Boh Kim is currently a professor in the Department of Computer Engineering, DaeJin University at Pocheon in Korea. He received the M.S degrees in Computer science from Yensei University in Seoul, Korea and the Ph.D degree in Computer Science from Catholic University of Daegu in Korea 1998 respectively. He spent two years as a Director of Central Computer Center in Korean Navy and also spend

three years as an assistant professor of Pukyong National University in Korea before joining the Daejin University. He was visiting professor at the University of Missouri at Rolla from 1999 to 2000. He was Head of computer Center in Daejin University. His research interests are in the following areas: software engineering, system analysis and design, and e-Biz software system.

Hyunsook Chung

Hyunsook Chung is a professor in the Department of Computer Engineering at Chosun University, Gwang-ju, Korea. She received a B.S. degree in Physics in Daegu Catholic University and a M.S. degree in Computer Engineering from same university. She received a Ph.D. degree in Computer Engineering from Yonsei University in 2003. Her research interests include multimedia data processing, knowledge data processing, ontology engineering, and information processing.