An Issue-oriented Syllabus Retrieval System based on Terminology-based Syllabus Structuring and Visualization

Hideki Mima[1]
(1) School of Engineering, University of Tokyo, Hongo 7-3-1, Bunkyo-ku, Tokyo 113-0033, Japan
mima@t-adm.t.u-tokyo.ac.jp n

ABSTRACT

The purpose of this research was to develop an issue-oriented syllabus retrieval system that combined terminological processing, information retrieval, similarity calculation-based document clustering, and visualization.

Recently, scientific knowledge has grown explosively because of rapid advancements that have occurred in academia and society. Because of this dramatic expansion of knowledge, learners and educators sometimes struggle to comprehend the overall aspects of syllabi. In addition, learners may find it difficult to discover appropriate courses of study from syllabi because of the increasing growth of interdisciplinary studies programs. We believe that an issue-oriented syllabus structure might be more efficient because it provides clear directions for users. In this paper, we introduce an issue-oriented automatic syllabus retrieval system that integrates automatic term recognition as an issue extraction, and similarity calculation as terminology-based document clustering. We use automatically-recognized terms to represent each lecture in clustering and visualization. Retrieved syllabi are automatically classified based on their included terms or issues. The main goal of syllabus retrieval and classification is the development of an issue-oriented syllabus retrieval website that will present users with distilled knowledge in a concise form. In comparison with conventional systems, simple keyword-based syllabus retrieval is based on the assumption that our methods can provide users, and, in particular, novice users (students), with efficient lecture retrieval from an enormous number of syllabi. The system is currently in practical use for issue-oriented syllabus retrieval and clustering for syllabi for the University of Tokyo's Open Course Ware and for the School/Department of Engineering. Usability evaluations based on questionnaires used to survey over 100 students revealed that our proposed system is sufficiently efficient at syllabus retrieval.

KEYWORDS: Issue oriented, syllabus retrieval, term extraction, knowledge structuring, visualization

Proceedings of the Workshop on Speech and Language Processing Tools in Education, pages 25–36,
COLING 2012, Mumbai, December 2012.

1 Introduction

Recently, scientific knowledge has grown explosively because of rapid advancements that have occurred in academia and society.[1] This rapid expansion of knowledge has made it increasingly difficult for learners and educators to comprehend the overall aspects of syllabi. In addition, because of the rapid growth of interdisciplinary studies programs, such as energy studies and earth-environmental studies, learners have found it increasingly difficult to discover appropriate courses of study in their syllabi.

Syllabus retrieval is believed to be one of several solutions to these problems. In fact, several syllabus retrieval systems have been proposed. In general, current syllabus retrieval methods can be classified as query-oriented and/or issue-oriented. Although the query-oriented method is useful and possesses strong retrieval capabilities, it can be difficult to employ, especially for novices, because the generation of queries usually requires users to first clarify their subjects.

The issue-oriented syllabus retrieval method was developed in an attempt to provide clear directions to learners. The issue-oriented syllabus structure is believed to be more efficient for learning and education, because it requires less knowledge about subjects (Mima et al., 2006). However, this system generally requires that users classify all syllabi manually in advance. This can be a time-consuming task. Thus, we can see that it is important to develop a more efficient method for automatic syllabus structuring to accelerate syllabus classification. The advantage of this technique is based on the assumption that automatic methods will enable more efficient processing of enormous amounts of syllabi texts.

In this paper, we introduce an innovative issue-oriented automatic syllabus classification system. We integrate automatic term recognition as issue extraction, terminology-based similarity-calculation for clustering, information retrieval, and visualization. Automatically-recognized terms are used to represent each lecture (or class) in clustering. In the system, provided syllabi are automatically classified and labeled according to the included terms that were automatically extracted. The main goal of syllabus retrieval and clustering is to develop an issue-oriented syllabus retrieval website that will present distilled knowledge to users in a concise form. The advantage of this system, in comparison with conventional syllabus retrieval or classification, is based on the assumption that automatic methods can efficiently process enormous amounts of text. The system has already been put into practical use for syllabus retrieval and clustering for the University of Tokyo's Open Course Ware and for the School/Department of Engineering syllabi. Usability evaluations based on questionnaires used to survey over 100 students revealed that our proposed system is sufficiently efficient at syllabus retrieval.

In the following section of this paper, we briefly explain the process of issue-oriented syllabi retrieval. We also provide an overview of the clustering system. In Section 2, we describe our proposed syllabus retrieval and classification scheme that is based on the use of automatically-extracted terms and on a visualization technique. In Sections 3 and 4, we discuss terminological processing as a feature extraction from each syllabus for similarity calculation and

1 For example, the Medline database (http://www.ncbi.nlm.nih.gov/pubmed) currently contains over 16 million paper abstracts in the domains of molecular biology, biomedicine, and medicine. The database is growing at a rate of more than 40,000 abstracts each month.

visualization. In Section 5, we present our evaluations of data collected from questionnaires used to survey over 100 students. We relied on the collected data to analyze the usability of our proposed scheme and to confirm its feasibility and efficiency. In the final Section, we present a summary of our approach and our conclusions.

2 System Overview

The main purpose of this study was to develop an efficient issue-oriented syllabus retrieval system that would provide clear directions to learners. Our approach to this issue-oriented syllabus classification system is based on the following:

- automatic term recognition (ATR) for automatic issue extraction
- automatic term clustering (ATC) for term variation management
- terminology-based document similarity calculation to develop syllabus classification
- automatic class label inference to clarify general issues of the classes

The system architecture is modular. It integrates the following components (see, Figure 1):

- *Terminology-based issue extraction (TIE)* – A component that conducts automatic term recognition as issue extraction from syllabus texts. It includes term extraction and term variation management.
- *Syllabus retriever (SR)* – It retrieves syllabi based on selected issues that are automatically extracted by TIE. It calculates similarities between each issue and each retrieved syllabus. Currently, we have adopted tf*idf based similarity calculation.
- *Similarity Calculation Engine(s) (SCE)* – It calculates similarities between KSs provided from each KR component by the use of ontology developed by ODE to show semantic similarities between each KSs. We adopted Vector Space Model-based (VSM) similarity calculation and we used terms as features of VSM. Semantic clusters of KSs were also provided.
- *SVM-based learning (SBL)* – A component that learns how to classify syllabi by extraction of classification patterns from features that have also been extracted by TFE. It then produces classification knowledge.
- *Terminology-based syllabus classification (SBC)* – It calculates similarities between syllabi provided by the SR component by the use of terms provided from TIE to develop clusters of syllabi. We adopted Vector Space Model-based (VSM) similarity calculation.
- *Term-based label inference (TLI)* – It infers representing labels for each class developed by TSC. We currently inferred labels based on term frequency (tf) for importance and document frequency (df) for generality.
- *Syllabus class visualizer (SCV)* – It visualizes syllabi structures based on graph expression in which classes of syllabi and representing labels of classes inferred by (TLI) are automatically provided.

As shown in Figure 1 and the flows by numbers, the system extracts issues automatically from syllabi texts in advance and produces classification of lectures based on these terms or issues. Then, representing labels (i.e., class labels) are also inferred by the use of terminological information. Finally, SVC visualizes syllabi structures with respect to selected issues.

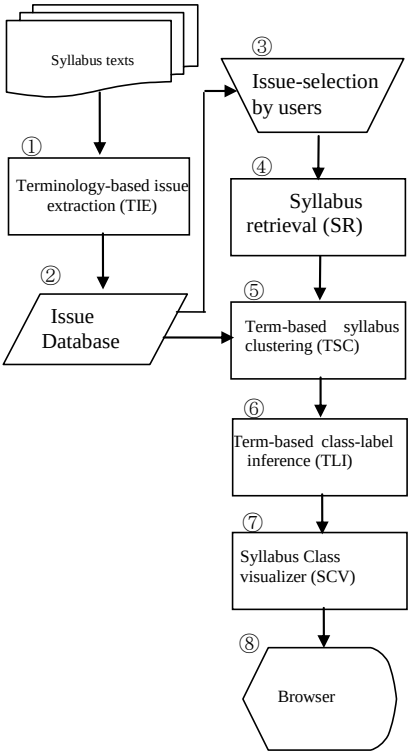

FIGURE 1 – The system diagram

3 Terminological processing as an ontology development

The lack of clear naming standards within a domain (e.g., biomedicine) makes ATR a non-trivial problem (Fukuda et al., 1998). Also, this lack of standards may typically cause many-to-many relationships between terms and concepts. In practice, two problems stem from this issue: (1) some terms may have multiple meanings (i.e., *term ambiguity*), and, conversely, (2) some terms may refer to the same concept (i.e., *term variation*). Generally, term ambiguity exerts negative effects on IE precision; term variation decreases IE recall. These problems reveal the difficulty involved in the use of simple keyword-based IE techniques. Therefore, the development of more sophisticated techniques, such as the identification of groups of different terms that refer to the same (or similar) concept(s) that could benefit from reliance on efficient and consistent ATR/ATC and term variation management methods, is needed. These methods are also important tools that can be used to organize domain-specific knowledge because terms should not be treated

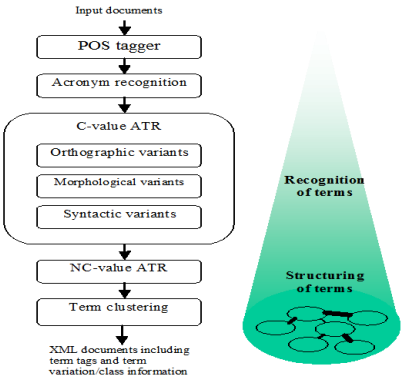

FIGURE 2 – Term recognition as issue extraction

in isolation from other terms. Rather, they should be related to one another so that relationships that exist between corresponding concepts are, at least partially, reflected in the terminology.

3.1 Term recognition

For our system, we used an ATR method based on *C/NC-value* methods (Mima et al., 2001; Mima and Ananiadou, 2001). The *C-valu*e method recognizes terms by combining linguistic knowledge and statistical analysis. The method extracts multi-word terms,[2] and it is not limited to a specific class of concepts. It is implemented as a two-step procedure. In the first step, term candidates are extracted by the use of a set of linguistic filters that describe general term formation patterns. In the second step, the term candidates are assigned termhood scores (referred to as C-*values*) based on a statistical measure. The measure amalgamates four numerical corpus-based characteristics of a candidate term: (1) frequency of occurrence, (2) frequency of occurrence as a substring of other candidate terms, (3) the number of candidate terms that contain the given candidate term as a substring, and (4) the number of words contained in the candidate term.

The *NC-value method* further improves the C-*value* results by considering the context of the candidate terms. The relevant context words are extracted and assigned weights based on the frequency with which they appear with top-ranked term candidates extracted by the *C-value* method. Subsequently, context factors are assigned to candidate terms according to their co-occurrence with top-ranked context words. Finally, new termhood estimations, referred to as *NC-values*, are calculated as a linear combination of the *C-values* and context factors for the respective terms. Evaluation of the *C/NC-methods* (Mima and Ananiadou, 2001) has revealed that contextual information improves term distribution in the extracted list because it places real terms closer to the top of the list.

[2] More than 85% of domain-specific terms are multi-word terms (Mima and Ananiadou, 2001).

3.2 Term variation management

Term variation and ambiguity have caused and continue to cause problems for ATR, as well as for human experts. Several methods for term variation management have been developed. For example, the BLAST system (Krauthammer et al., 2000) used approximate text string matching techniques and dictionaries to recognize spelling variations in gene and protein names. FASTR (Jacquemin, 2001) handles morphological and syntactic variations by means of meta-rules used to describe term normalization. Semantic variants are handled via WordNet.

The basic *C-value* method has been enhanced by term variation management (Mima and Ananiadou, 2001). We consider a variety of sources from which term variation problems originate. In particular, we deal with orthographical, morphological, syntactic, lexico-semantic, and pragmatic phenomena. Our approach to term variation management is based on term normalization as an integral part of the ATR process. Term variants (i.e., synonymous terms) are addressed in the initial phase of ATR when term candidates are singled out. This differs from the process that is used in other approaches (e.g., FASTR handles variants subsequently by application of transformation rules to extracted terms). Each term variant is normalized (see, Table 1, as an example) and term variants that have the same normalized form are then grouped into classes to link each term candidate to all of its variants. In this way, a list of normalized term candidate classes, rather than a list of single terms, is statistically processed. The termhood is then calculated for a whole class of term variants, rather than for each term variant separately.

Term variants	Normalized term
human cancers cancer in humans human's cancer human carcinoma	} → human cancer

TABLE 1 – Automatic term normalization

3.3 Term clustering

In addition to term recognition, term clustering is an indispensable component of the literature mining process. Because terminological opacity and polysemy are very common in molecular biology and biomedicine, term clustering is essential for the semantic integration of terms, the construction of domain ontologies, and for semantic tagging.

In our system, ATC is performed by the use of a hierarchical clustering method in which clusters are merged based on average mutual information that measures the strength of the relationships between terms (Ushioda, 1996). The system uses terms automatically recognized by the *NC-value* method and their co-occurrences as input. A dendrogram of terms is produced as output. Parallel symmetric processing is used for high-speed clustering. The calculated term cluster information is encoded and used for calculation of semantic similarities in the SCE component. More precisely, the similarity between two individual terms is determined based on their position in a dendrogram. In addition, a commonality measure is defined as the number of shared ancestors between two terms in the dendrogram. A positional measure is defined as the

sum of their distances from the root. Similarity between two terms corresponds to a ratio between commonality and positional measure.

Table 3 shows a sample of automatically-recognized terms (issues) that occur in an Engineering domain syllabus text that consists of 850 lectures (Faculty of Engineering, University of Tokyo, 2006). As we can see from the Table, reasonable and representative issues were successfully extracted by our method.

Automatically-Recognized Terms	Termhood
基礎知識 (basic knowledge)	144.55
線形代数 (linear algebra)	77.35
統計力学 (statistical mechanics)	74.00
固体物理 (solid-state physics)	67.20
ベクトル解析 (vector calculus)	65.01
偏微分方程式 (partial differential equation)	62.40
材料力学 (mechanics of materials)	62.13
環境問題 (environmental issues)	60.17

TABLE 2 – Sample of recognized issues

Further details of the methods and their evaluations can be found in Mima et al. (2001) and Mima and Ananiadou (2001).

4 The Use of Visualization to Generate Issue-oriented Syllabus Structures

In our system, the TSC, TLI, and SCV are implemented by the integration of terminology-based issue extraction from syllabi and by clustering of syllabi based on semantic similarities that are also calculated based on terms in syllabi. Graph-based visualization for the automatic generation of issue-oriented syllabus structures is also provided to help in retrieval of lectures. Figure 3 shows an example of the visualization of issue-oriented syllabus structures relevant to the issue, "environment and energy," that occurs in the engineering syllabus. To structure knowledge, the system constructs a graph in which the nodes are used to indicate relevant syllabi for the key issues selected by the user. Links among the syllabi indicate semantic similarities that are calculated by the use of terminological information developed by our TIE components. Semantic similarity is based on comparisons of terminological information extracted from each syllabus, whereas conventional similarity calculation is generally based on extracted nouns. In addition, the locations of each node are calculated and optimized when the graph is drawn. The distance between nodes depends on the closeness of their meanings. The complete algorithm of this issue-structuring method is presented below:

begin

$Q \leftarrow$ issues specified to IR

$R \leftarrow$ IR(Q) // retrieving relevant syllabi to Q and putting them into R

for every x ***in*** R ***do***

w(Q, x) $\leftarrow$ *IRscore(Q, x) // calculate IR score between Q and x*

for every y ***in*** R ***do***

if $x \neq y$ ***then***

$p \leftarrow Ont(x)$ // retrieving terminological information of x

$q \leftarrow Ont(y)$ // ″ y

$w(x,y) \leftarrow Sim(p,q)$ // calculate similarity using p and q

fi

end

end

Visualize graph based on every $\{w(i,j)|i=Q \text{ or } i \in R, j \in R, i \neq j\}$

end.

We generate an issue-oriented syllabus structure based on (1) cluster recognition and (2) terminology-based cluster label inference. Cluster recognition is performed by detection of groups of nodes in which every combination of included nodes is strongly linked (i.e., their similarity exceeds a threshold). Automatic cluster label inference is performed by the use of terminological information included in each cluster with respect to tf (term frequency) and df (document frequency (i.e., term generality)).

5 Evaluation

We performed a practical application of the system for syllabus retrieval for the University of Tokyo's Online Course Catalogue (UTOCC),[3] for the Open Course Ware (UT-OCW)[4] site, and for the syllabus-structuring (SS) site[5] for the School/Department of Engineering. All of these syllabi are available to the public over the Internet. The UT-OCW's course search system is designed to search the syllabi of courses posted on the UT-OCW site and on the Massachusetts Institute of Technology's OCW site (MIT-OCW). In addition, OCC and SS site's search is designed to search the syllabi of more than 9,000 lectures from all schools/departments at the University of Tokyo, and 1,600 lectures from the School/Department of Engineering at the University of Tokyo. Both systems display search results based on relationships that exist among the syllabi as a structural graphic (see, Figure 3). Based on terms that were automatically-extracted terms (issues) from the syllabi and on similarities calculated by the use of those terms, the system displays the search results in a network format that uses dots and lines. In other words,

[3] http://catalog.he.u-tokyo.ac.jp/
[4] http://ocw.u-tokyo.ac.jp/.
[5] http://ciee.t.u-tokyo.ac.jp/.

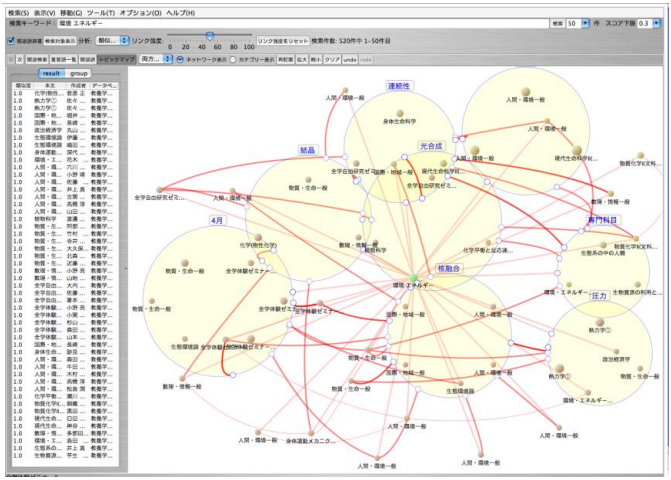

FIGURE 3 – Issue-oriented syllabus structuring: Visualization sample

the system extracts issues from the listed syllabi. It rearranges these syllabi based on semantic relationships that occur in the contents. It displays the results graphically. This differs from conventional search engines that simply list syllabi related to keywords. Because of this process, we believe users will be able to search for key information and obtain results in a minimal amount of time. In graphic displays, as mentioned previously, relevant syllabi are shown in a structural graphic that uses dots and lines with cluster circles. Stronger semantic relationships that occur in the syllabi or clusters will be located in closer proximity on the graphic. This structure will enable users to find groups of courses/lectures that are closely related to some issues. It will also help users take courses/lectures in logical order (e.g., a user can begin with fundamental mathematics and proceed to applied mathematics). Furthermore, if they consult the structural graphic display, users will be able to instinctively find relationships among syllabi drawn from other faculties or universities.

Currently, we have obtained over 2,000 hits per day, on average, from sites worldwide. We have provided more than 50,000 page views during the last three months.

We conducted a usability evaluation based on questionnaires used to survey 120 novice students. We obtained positive statements about the efficiency of syllabus retrieval by the search system from more than 70% of respondents.

Finally, we obtained 151 positive statements and 168 statements that recommended further improvement. Tables 3 and 4 demonstrate the breakdown for positive statements and the breakdown for statements that recommended further improvement, respectively. As can be seen in Table 3, we can reasonably state that our issue-oriented scheme and the related system is relatively efficient at syllabus retrieval. On the other hand, as can be seen in Table 4, we must continue to make further improvements. In particular, we must improve the visualization scheme and its scalability to link other syllabi DB and systems.

Positive statements	#
Advantage of visualization	45
Improvement in retrieval efficiency	41
Clarity of results	22
User-friendly interfaces	20
Misc.	23
Total	151

TABLE 3 – Breakdown of positive statements

Statements that recommended further improvement	#
Complexity of visualization	67
Additional linkage to other syllabi	23
Lack of clarity about relationships that exist among lectures	11
Linkage to other systems (e.g., lecture management, etc.)	13
Quality of issue extraction	10
Difficulty of operation	5
Speed of calculation	1
Misc.	38
Total	168

TABLE 4 – Statements that recommend further improvement

Conclusion

We developed an issue-oriented syllabus retrieval system that combined terminological processing, information retrieval, similarity calculation-based document clustering, and

visualization. The system provides visualizations of issue-oriented syllabus structuring during retrieval. This differs from conventional syllabus retrieval that solely provides a list of retrieved results relevant to a specific query.

We evaluated the system based on data collected from questionnaires used to survey over 100 students. Based on our results, we can reasonably state that the system provides relatively efficient syllabus retrieval.

References

Fukuda, K., Tsunoda, T., Tamura, A. and Takagi, T. (1998). *Toward information extraction: Identifying protein names from biological papers*, Proc. of PSB-98, Hawaii, pp. 3:705–716.

Mima, H., Ananiadou, S. and Matsushima, K. (2006). *Terminology-based Knowledge Mining for New Knowledge Discovery*, *ACM Transactions on Asian Language Information Processing* (TALIP), Vol. 5(1), pp. 74–88.

Mima, H., Ananiadou, S. and Nenadic, G. (2001). *ATRACT workbench: An automatic term recognition and clustering of terms.* In V. Matoušek, P. Mautner, R. Mouček, K. Taušer (eds.) Text, Speech and Dialogue, LNAI 2166, Springer Verlag, pp. 126–133.

Mima, H. and Ananiadou, S. (2001). *An application and evaluation of the C/NC-value approach for the automatic term recognition of multi-word units in Japanese*, *International Journal of Terminology*, Vol. 6(2), pp. 175–194.

Krauthammer, M., Rzhetsky, A., Morozov, P. and Friedman, C. (2000). *Using BLAST for identifying gene and protein names in journal articles. Gene* 259, pp. 245–252.

Jacquemin, C. (2001). *Spotting and discovering terms through NLP*. MIT Press, Cambridge MA, p. 378.

Ushioda, A. (1996). *Hierarchical clustering of words.* In Proc. of COLING '96, Copenhagen, Denmark, pp. 1159–1162.