A Context-based Rating System for Online Communities

by

Eugene Chiu

Submitted to the Department of Electrical Engineering and Computer Science

in Partial Fulfillment of the Requirements for the Degrees of

Bachelor of Science in Electrical Engineering and Computer Science

and Master of Engineering in Electrical Engineering and Computer Science

at the Massachusetts Institute of Technology

May 24, 2002
[June 2002]

Author____
Department of Electrical Engineering and Computer Science
May 24, 2002

Certified by____
Peter Szolovits
Thesis Supervisor

Accepted by ____
Arthur C. Smith
Chairman, Department Committee on Graduate Theses

A Context-based Rating System for Online Communities
by
Eugene Chiu

Submitted to the
Department of Electrical Engineering and Computer Science

May 24, 2002

In Partial Fulfillment of the Requirements for the Degree of
Bachelor of Science in Electrical Engineering and Computer Science
and Master of Engineering in Electrical Engineering and Computer Science

ABSTRACT

In today's Internet world, a myriad of online communities have sprung up to address our every need, whether it be for auctioning used goods, reading news, or learning about photography. With so much information on a site, it would be useful for users to be able to represent their interests in a profile so that content and ratings would be tailored to their preferences. This document provides a system for personalizing an online community using a fixed ontology and Internet technologies. Redbook, a resource-finding Web site for the Harvard-MIT medical community, was implemented and evaluated for this project. We will discuss the methods and algorithms used for constructing the personalization features for Redbook and then analyze the success of this endeavor based on testing in an artificial environment.

Thesis Supervisor: Peter Szolovits
Title: Professor of Computer Science and Engineering

Table of Contents

CHAPTER 1

INTRODUCTION

Given the number of users and amount of content on an online community, it is often difficult to determine what information on the site is useful and relevant to an individual user. Many sites incorporate a ratings system for users to judge site content so that the content quality can be assessed; however in most cases, the ratings system is based on an average evaluation by all other raters, with no regard to the expertise or interests of the rater or his relationship to the subject matter of the documents being rated. It will be valuable for a site to incorporate a system where users can effectively describe their interests and expertise in a standardized manner, and use this classification to gauge similarities either between the interests of users and content or between users and raters. Through such a system, the similarities between users and content can be leveraged to personalize the ratings system on the site so that both content delivery and the ratings applied to content can more accurately reflect a viewer's interests and priorities.

The following thesis describes the design and implementation of a personalized rating system in an online collaborative environment called the Redbook. In this introduction I will briefly discuss background and motivations behind the project and other efforts which have been made in this research area. I will also describe the nature of the Redbook environment and the goals for implementing the ratings system into this environment.

1.1. Background and Motivations

Next to its users, one of the most important components of an online community is content, which might include posted messages, online tutorials and software tools, or information and news links. Since membership in these online communities can be quite diverse (potentially open to anyone with Internet access in the most public communities), submitted content can vary greatly in quality and relevance, ranging from the highly informed content submitted by domain experts, to non-informative or incorrect data submitted by a novice. Although it is theoretically possible for a fixed group of administrators to constantly monitor the site and judge content, this can consume large amounts of time and monetary resources. Given that many members of the online community are themselves highly expert in judging the quality of submitted content, it may be more feasible and useful to have the users themselves rate content on an ongoing basis.

The notion of user-derived ratings of content is not new. Besides rating content there are sites whose member-users rate products and services and offer appraisals of the reputation or reliability of other members of the user-group. The commercial auction site eBay, for example, relies heavily on member-derived ratings to establish the reputation of other eBay community members.[1] Knowing how well the larger community knows and trusts a particular seller is thought to be a valuable tool in reassuring a potential customer prior to executing a transaction.

Among online communities that use a rating system, most use a simplified model in which all raters have equal credibility and there is no differential given to ratings

[1] www.ebay.com

derived from domain experts versus novices. Depending on the nature of the site, there are cases where it would be very beneficial to incorporate a weighted ratings scheme, one that takes into account the user's background and expertise. In this manner, heavier weightings could be applied to ratings from users with similar backgrounds (and thus presumably similar preferences) and users with expertise in the content being rated (and thus greater authority). The site would display personalized ratings to each individual user tailored to the user's preferences and background. This triangular process is illustrated in its simplest case below, when there is one viewer, one rater, and content.

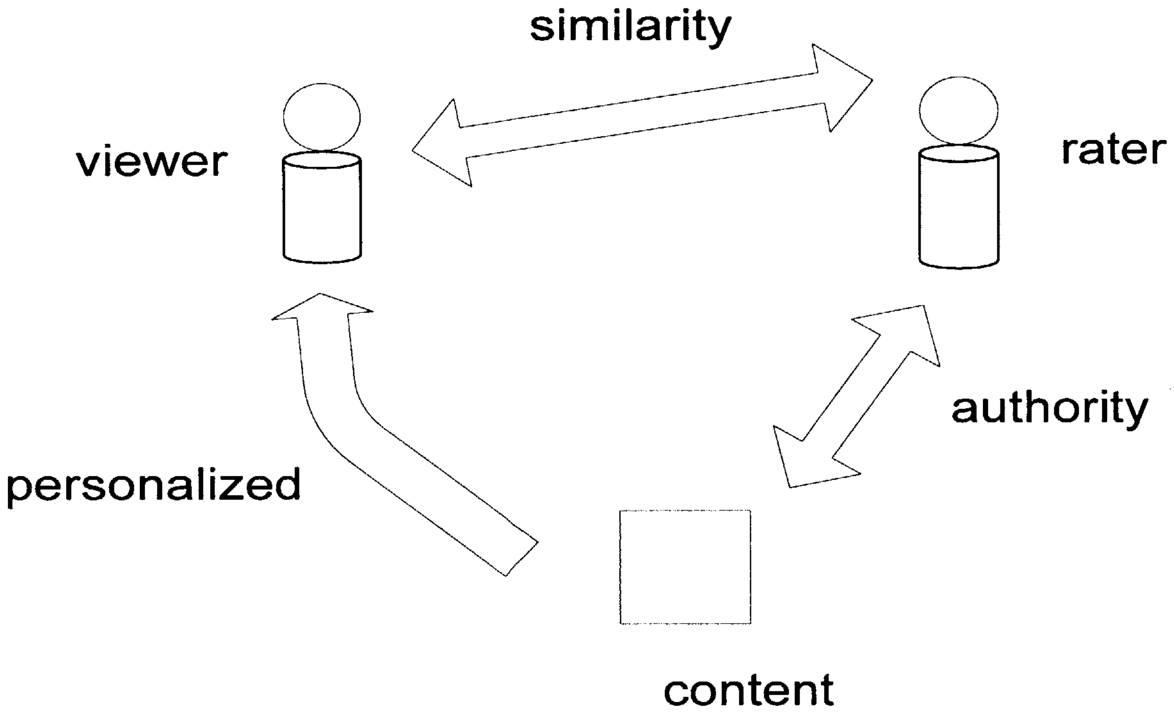


Fig. 1- Triangle relationship- A viewer looking at content on a site sees a personalized rating based on how similar he is to the raters and how much authority the raters have pertaining to the content.

1.2. Related Web Sites

Slashdot, an online site for posting technical news[2], has a system of moderation and "karma points" which allows those users who are perceived to have good reputations to rate posted comments. The higher a user's "karma," the more likely it is that he will be allowed to moderate, or rate content. This weighting scheme is counterbalanced by a second facility called meta-moderation, in which the consistency or quality of a particular moderator's performance is periodically assessed by other users. Presumably, this encourages moderators to be honest in their assessment of content. When users are granted the ability to moderate or meta-moderate, they have a limited number of uses based on the number of points they have. This prevents a bad moderator from wreaking havoc across the site through intentionally inappropriate ratings[3].

Another online community that has recently developed, www.Devhood.com, enables college students to share information about Microsoft Corporation's .NET technologies. The content of this site includes tutorials, message boards with message archives, new software tools, and news links. The tutorials and software/development tools may be rated on their quality and pertinence to particular tasks. In an effort to make such ratings more useful and reliable, Devhood incorporates a "caste system" to stratify users according to the quality and frequency with which they have posted and reviewed content. In this system, points are awarded to users based on the amount of content they have posted and the quality of that content. Over time, with repeated high quality contribution and critiquing, the user may accumulate a large number of points which translate into a higher status in the "caste system." The more a user participates, the more

[2] www.slashdot.org

[3] www.slashdot.org

experience points he has. Content ratings are subsequently weighted based on the experience of the rater; higher level users will have more clout than new users. There is no limit to how much content a user can rate, except that they can only rate each item once.[4]

1.3. Redbook

The original Redbook was a booklet distributed in the Harvard-MIT Division of Health Sciences and Technology that sought to provide incoming students with descriptions of research opportunities, so that they could make informed decisions of which groups they could join. We are currently in the process of creating an online version of Redbook that will provide students and researchers with information and advice necessary to make informed decisions about a range of research-related topics. In addition to containing descriptions of active and future topics by research groups, the Redbook environment might include funding opportunities, research tools, established or potential collaborations, as well as recommendations and critiques of research settings, research groups, projects, seminars, collaborations, and personal research experiences. The information will be updated on a regular basis rather than being distributed annually.

Redbook users are current and future students, researchers (PIs, post-doctoral fellows, research associates, potential collaborators) and clinicians affiliated with the Massachusetts Institute of Technology and Harvard University programs in biological and medical sciences. Student users of the Redbook search for and attempt to identify high quality research opportunities that may serve as a basis for their doctoral degree. Researchers use the Redbook to characterize themselves publicly in terms of research

[4] www.devhood.com

interests and activities in an effort to attract talent. Redbook users thus represent members of a disseminated community who may be in search of each other but are not certain of how to best identify, locate and connect. Ultimately, the quality of the ratings that are applied to the resources will determine the success or utility of Redbook to students and researchers, alike. Content descriptions and ratings that are perceived as valid and reliable will result in increased use of the system and potentially improve the process by which students are matched to research experiences.

Profiles will be established on the Redbook site in order to acquire information about the user, information which would be helpful in weighing ratings. There are different kinds of profiles which might be useful, including research profiles, academic profiles, volunteerism profiles, and political profiles. Each one of these provides a separate body of information about the user which can be applied appropriately depending on the nature of the content being rated. Profiles are useful because they can be edited and updated to reflect any changes in a user's experiences and interests. The design should be such that there is a way to quantify similarities between profiles.

A Redbook profile will be very important in describing a user's interests for any one of a variety of purposes. Most notably, the profile will be used to accurately reflect a user's research interests so that relevant content, such as course descriptions, event announcements or topic-specific documents can be directed to the user. An ontology (Section 2.2) has been built to assist users in properly describing their interests with terms and concepts derived from the ontology. In addition to allowing users to specify their research interests, the profile also allows a user to rate what "aspects" of an event he considers important. In Redbook, an aspect is one of many categories like time value, presentation, and organization which pertain to different qualities of an event. Since users value these qualities differently, it will be important for a user to be able to specify in his profile which aspects of an event he considers more important. Both the user's research interests and his preferences for event aspects are used in personalizing an event rating.

It is important for Redbook to not only allow users to create a profile, but to automatically and intelligently characterize other site content (such as events announcements, for example) based on their titles and descriptions. Since the same ontology is used to create profiles of both users and content, the classification of all Redbook entities is more uniform. This standardized classification scheme serves as a foundation by which similarity and relevance measurements can be applied to users and content. Based on similarity measurements, the ratings displayed to individual users can be dynamically adjusted to reflect the opinions of like-minded raters. The profiles also support rank-ordering content based on relevance to the user's stated preferences. In order to develop this system, the following features must be implemented for Redbook:

- Ontology-assisted search and profile creation for users
- Automated profile creation for content
- Relevance determination between profiles

The development of these capabilities, described in Chapter 3, allows for the creation of a personalized ratings system on Redbook, which is the ultimate goal of my project. Chapter 4 goes into further detail about how event ratings will be personalized based on a user's research interests and preferences for event aspects.

CHAPTER 2

ONTOLOGIES

As stated in the introduction, the main features to be implemented on Redbook include an ontology-assisted search and profile creation for users, automated profile creation for content, and relevance determination between profiles. Since all of these features rely on a standardized classification scheme, I will begin with a discussion of the ontology that will support the system.

2.1. Objective

The World Wide Web is by far the largest source of shared information on the planet; there is so much information on the web and no standard way to represent the information, since the Internet receives contributions from a myriad of people with different backgrounds, interests, and skill levels. There has been much research done in finding ways to effectively search the World Wide Web and retrieve the desired information. The three main categories of search mechanisms include:

- Text-indexing engines
- Hand-build catalogs
- Private robots using methods to gather limited semantic information

Text-indexing is at a disadvantage because it uses strict syntactic and lexical matches, so in many cases information which is related is ignored. Hand-built catalogs require too many man-hours to create to be of any use, and robots gathering information is currently ineffective because natural language processing still is at its early stages.[5]

[5] Luke S, Spector L, Rager D, Hendler J. "Ontology-based Web Agents." (University of Maryland and Hampshire College, 1996.)

The Redbook site looks to be in a similar situation as the Internet except on a much smaller scale; it will receive contributions like event notifications and document tutorials from students and faculty all over the Boston area. Instead of relying on administrators to standardize all the characterizations of site content, the Redbook site will need a system to automatically characterize the users and content and go beyond the normal search by keyword functionality which is prevalent on the Web. Keyword searching usually does not return all of the desired results, because the search query is too narrow in scope to be of much use to the user. On Redbook in particular, users searching for project groups or events may only have a vague idea of what their research interests are. If a tool could be provided to help users expand their search by providing conceptually related keywords, it would make resource-finding on Redbook a much easier and effective process.

2.2. Solution - Ontology System

In recent years ontologies have been developed and introduced into computer-science applications to help resolve this issue of searching a large body of textual data. It is difficult to come up with a formal definition of an ontology, but it has been characterized in the following quote:

> "An ontology may take a variety of forms, but necessarily it will include a *vocabulary of terms*, and some *specification of their meaning*. This includes definitions and an indication of how concepts are inter-related which collectively impose a structure on the domain and constrain the possible interpretations of terms."[6]

An ontology is effective because it can help establish equivalence in meaning to information that is represented in a lexically variable manner. It can also expand upon the meaning of a set of terms or provide mappings between distinct, but semantically-related concepts[7]. In the case of the Redbook site an ontology can assist in the matching of people to content that interests them by drawing matches beyond the keywords which are specified in content titles and descriptions. By simply functioning as a taxonomical tree without any specifications of meaning, the Redbook ontology can accomplish the goal of assisting users in resource-finding. The ontology helps users characterize concepts using standardized terminology and precise relationships that may exist between different terms or phrases.

[6] Jasper R and Uschold M. "A Framework for Understanding and Classifying Ontology Applications." (Boeing Math and Computing Technology, 1999.)

[7] Guarino, Nicola. "Formal Ontology and Information Systems." (National Research Council, 1998.)

2.3. Constructing Ontologies

The first step in developing an ontology is to define the extent of the domain that is being modeled. One of the approaches[8] which is appropriate for Redbook purposes starts out by defining the purpose of the ontology, which in this case is information modeling. More specifically for Redbook, the ontology will model research concepts dealing with research opportunities at MIT and the Harvard Medical community

The second step to building an ontology deals with the actual modeling of concepts in the specific domain, and it is composed of two main activities:

- Ontology Capture
- Integration of Existing Ontologies

Ontology capture refers to the identification and capture of important concepts and their relations in order to lay out the framework of the ontology and ensure that it will fully represent most if not all of the knowledge in the domain of interest. Explicit relationships that may exist between the concepts in the ontology are established in this activity. Integration of existing ontologies is also a very important activity and will assist greatly in the development of an ontology by both reducing the modeling effort and acquiring variety in the knowledge representation.

There are many tools available which will assist in the construction of an ontology. The Knowledge Systems Laboratory at Stanford University has developed a software tool called "Ontolingua" which provides a collaborative environment for users

[8] Stubkjaer, E. "Integrating Ontologies: Assessing the Use of the Cyc Ontology for Cadastral Applications." (Aalborg University, 2000.)

to create, edit, modify, and use ontologies.[9] Many people have set up their projects on the Ontolingua server, and the services are also accessible over the web through a standard browser.

Protégé is a software tool developed by the Stanford Medical Informatics group which also allows users to create and modify ontologies. It provides multiple graphical extensions which enable users to access other knowledge-based applications and contains a library which the other knowledge-based systems can access. Protégé's main purpose is to provide a platform for knowledge-based editing, while Ontolingua is more focused on the sharing and collaboration of knowledge across the Web.

[9] http://www.ksl.stanford.edu/software/ontolingua/

2.4. Redbook Ontology

For the Redbook site, the ontology will initially be limited to a hierarchical format, with relationships limited to node-subnode format.

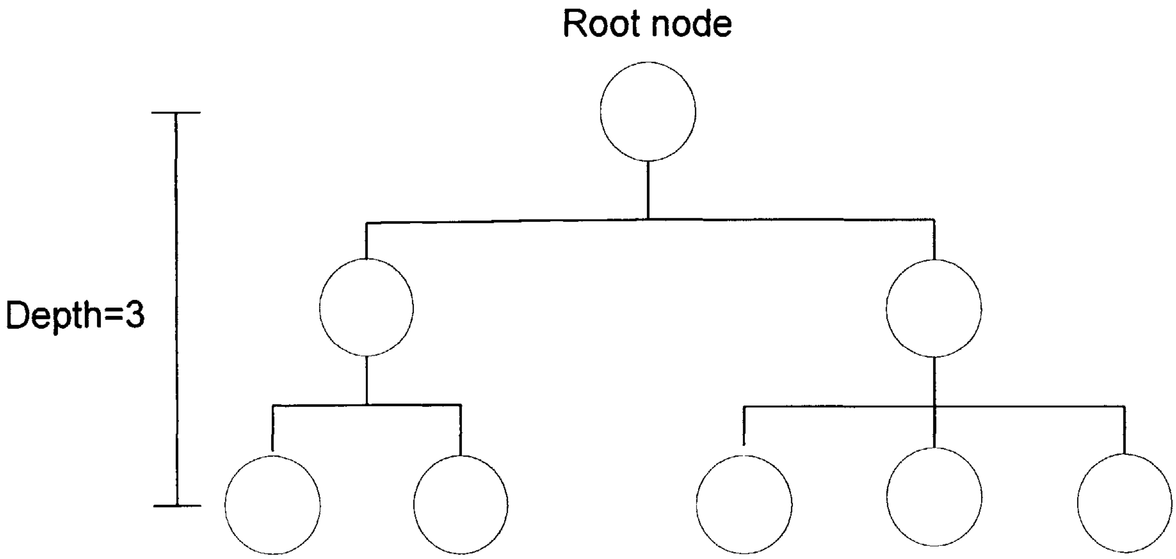


Fig. 2- Tree-structure ontology

Starting at a root node, the Redbook ontology expands into a series of successively branching child nodes. Called a broader-narrower hierarchy, the tree structure of the Redbook óntology starts off with broad terms and goes into more specific terms as the nodes branch further down. The depth of any specific node refers to its distance from the root, and as a reference the depth of the root itself is one. Each node besides the root node has exactly one parent node from which it branches, and all the children of a single parent are considered peer nodes of each other.

To support the Redbook's primary goal of identifying biomedical research resources and matching them to users, the ontology was developed using the hierarchical scheme to represent concepts from the scientific and medical domains. With respect to the medical domain, some conceptual structures were derived or modified from the

National Library of Medicine's Medical Subject Heading (MeSH) scheme. Since a large volume of the biomedical literature is indexed using this MeSH scheme, it seemed appropriate to adopt some of this structure to facilitate characterization of recent research activities by these authors. The final ontology constructed[10] for Redbook contained two-thousand nodes and had a maximum depth of ten.

The use of ontologies in Redbook satisfies two major needs of a resource discovery environment: the first need is for a more efficient way for a user to search the database for projects or other documents of interest to them without over-reliance on specific string matching, and the second need is a metric to measure relevance or similarity between entities. Here, entities will be defined as either a user or a specific resource. Thus, similarity may be measured in terms of two different users who share specific research interests, or a single user's interests in relation to the content of a resource.

2.4.1. Central versus Individual Perception

Ontologies strive to establish a consensus or canonical representation of a specific body of knowledge in terms of a standardized vocabulary. Clearly, no single ontology is capable of capturing the variations in conceptual structures that exist between different users. Acknowledging that our ontology may still over-represent or under-represent various concepts, we have attempted to tailor the ontology to conform to the conceptual structure shared by its target entities. For our audience of users, we made several assumptions. First, in characterizing biology and engineering concepts, we assumed that student users would likely adopt a conceptual structure similar to that on which the

[10] Dierks, Meghan.

curriculum and coursework are based. Thus, we used extant descriptions of courses and lectures offered at MIT and taxonomies developed by bioengineering domain experts (originally developed to describe the curriculum at 4 major educational institutions - Vanderbilt University, Massachusetts Institute of Technology, Northwestern University and University of Texas) as a foundation for this portion of the ontology. Second, in characterizing medical concepts, we assumed that using a schema similar to that used to index most of the medical publications would result in good representation of publications in this domain. Hence, we used a subset of the MeSH concept hierarchy as a foundation for this part of the ontology. As long as the Redbook ontology can be accepted by a high majority of users looking for resources such as those described above, it can be considered a good knowledge representation.

CHAPTER 3

ARCHITECTURE

There are three main features to be implemented on Redbook, the ontology-assisted search and profile creation, automated profile creation for content, and relevance determination between profiles. The following sections detail the implementation of these three features which will serve as the architectural framework for Redbook and allow for the creation of a personalized ratings system.

3.1. Ontology-assisted Search and Profile Creation for Users

In Section 2.4 it was stated that one of the two major needs satisfied by an ontology in Redbook was that an ontology provides for a more efficient way for a user to search the database without over-reliance on specific string-matching. Both the search and the profile creation processes in Redbook utilize the ontology in the same way to allow users to choose keywords which are related to their query. The only difference is that in creating a profile the user submits all the words found with the assistance of the ontology, while in searching the database the words are used to do a full-text search on the database. The components for doing a full-text search are introduced and discussed before detailing the implementation of ontology-assisted search and profile creation.

3.1.1. Similarity Search

In order to implement an ontology-assisted search, first a mechanism for searching the documents is needed. For finding matches based on keywords specified by a user as in most search engines, a similarity search is most often used. This approach represents the documents as an inverted index, which is a data structure containing an entry for each term. Each entry contains a list of document identifiers for all the documents which contain an instance of the term; in addition, meta-information like word-frequency, position, or document-length which would help with calculating similarity might be stored along with each document identifier[11].

A limitation of a basic inverted representation for text indexing is that there is no way to look at the context of the document; rather relevance is calculated using things like word frequency and document length. In larger documents, this would be even more of a problem because huge bodies of text tend to have many words unrelated to the topic of the document. Latent Semantic Indexing solves this problem by mapping documents to a concept space and thus helping to eliminate noise from unrelated words to the search query. In order to allow indexing on these conceptual searches, conceptual word-chain representations of documents have been developed[12].

The inverted representation seems to be the method of choice in most applications. Based on its wide acceptance and convenience, the basic inverted representation was used to do a similarity search in Redbook documents. The noise effects from unrelated words will be minimal because titles and descriptions of the targeted resources are usually less than three-hundred characters. Also, the documents are stored in a Microsoft SQL Server database which enables full-text indexing on tables.

3.1.2. Microsoft Full-Text Search

All of the content tables in the Redbook database have full-text indexing enabled. Instead of scanning through all table columns in the database every time a query is

[11] Salton G, McGill MJ. "Introduction to Modern Information Retrieval." (McGraw Hill, 1983.)

[12] Aggarwal C, Yu P. "On Effective Conceptual Indexing and Similarity Search in Text Data." (IBM T.J. Watson Research Center, 2001.)

presented, much of the work is done beforehand. An indexing engine stores all the terms in a manner which makes it easy to retrieve. The data structure for storage is the previously described inverted index, which contains a row for each term along with information about the documents in which the term appears and the number of occurrences and relative position of the term within each document. These numbers provide the ability to apply formulas and probabilities during each search to determine the relevance of documents. Full-text search offers capabilities like ranking query results and doing stem-searches where the root form of the query term is expanded to include alternate forms.[13]

3.1.3. Ontology-assisted Search and Profile Creation

Redbook provides a full-text search for users to find appropriate projects matching their interests which they provide in a search query. This is complemented with an ontology using the following model:

[13] "Microsoft Full-Text Search Technologies: White Paper." (Microsoft Corporation, 2001.)

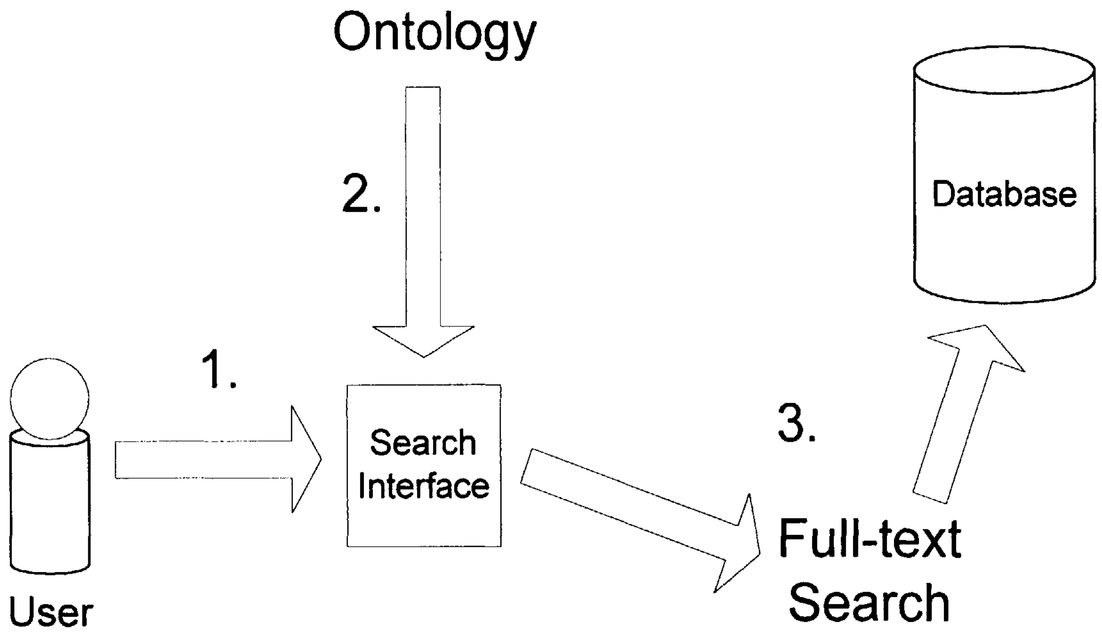
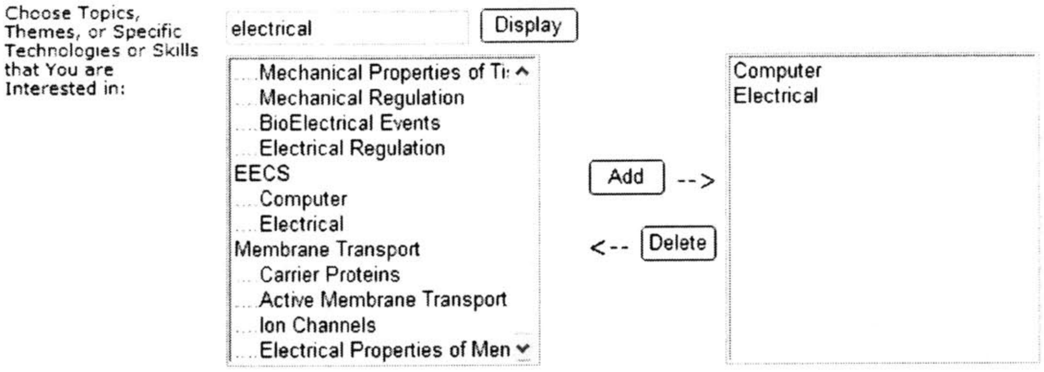


Fig.3- Ontology Search Model

The ontology is created in order to assist users in identifying concepts which they are interested in, based on the step-by-step process described below and illustrated above in Figure 3:

1. The user is provided with a textbox to enter in a term which they are looking for.
2. The term is sent to the ontology, and any concept nodes in the ontology which contain the term are returned along with all the parents, children, and peers of those nodes.
3. From this list of concepts the user chooses any additional terms which are of interest and submits these terms as a final search query to the full-text enabled database.

The presentation of related concepts from the ontology allows a user to expand his original search query to include related terms which might lead to the discovery of projects of interest, or incrementally refine his search to more specific terms.



Choose Topics, Themes, or Specific Technologies or Skills that You are Interested in: electrical

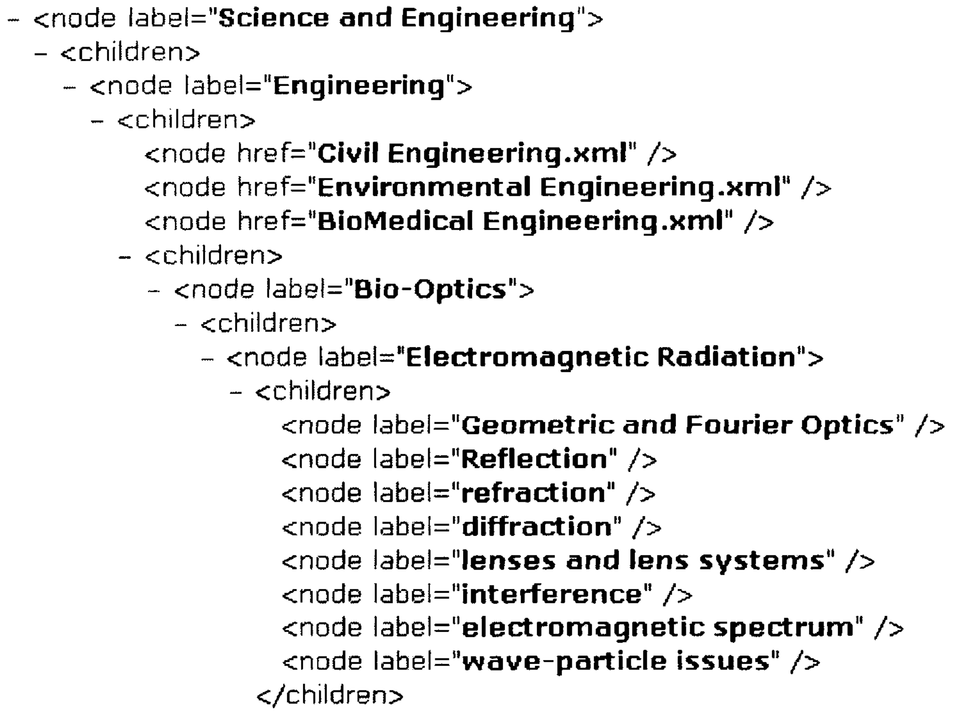
Computer
Electrical

Display
....Mechanical Properties of Ti:
....Mechanical Regulation
....BioElectrical Events
....Electrical Regulation
EECS
....Computer
....Electrical
Membrane Transport
....Carrier Proteins
....Active Membrane Transport
....Ion Channels
....Electrical Properties of Men
Add -->
<-- Delete

Fig.4- Search Interface- The user enters his search query "electrical" in the text box and clicks the "Display" button, which produces a list of all nodes in the ontology which contain the term along with all the parents, children, and peers of the nodes. In addition to selecting the "Electrical" concept node, the user also selects "Computer" and adds it to his final selection list box.

Profile creation for users is very similar to the process described above for the search, with the only difference being that the keywords obtained and added into the final selection list box are stored in the user's profile instead of being submitted as a full-text search query to the database. However, the effect of the ontology in finding additional conceptually related terms for the user was the same for both profile creation and search.

The ontology is implemented in an XML document and traversed using the <children> and <node> tags. The structure of the ontology tree in XML is shown below:



```
- <node label="Science and Engineering">
  - <children>
    - <node label="Engineering">
      - <children>
          <node href="Civil Engineering.xml" />
          <node href="Environmental Engineering.xml" />
          <node href="BioMedical Engineering.xml" />
        - <children>
          - <node label="Bio-Optics">
            - <children>
              - <node label="Electromagnetic Radiation">
                - <children>
                    <node label="Geometric and Fourier Optics" />
                    <node label="Reflection" />
                    <node label="refraction" />
                    <node label="diffraction" />
                    <node label="lenses and lens systems" />
                    <node label="interference" />
                    <node label="electromagnetic spectrum" />
                    <node label="wave-particle issues" />
                  </children>
```

Fig. 5- XML Schema for Redbook Ontology- A sample section of the Redbook Ontology

At each level the node is described by its label, and if it has children then there are nodes underneath it enclosed by the <children> tags. The system traverses the tree by inspecting each node's label to see if it matches the one it's looking for, then recursively searches the children nodes. Using this XML Schema any tree which is appropriate for Redbook can be represented; there are no cyclic trees used for the ontology.

3.2. Automated Profile Creation for Content

One of the roles of the Redbook is to automatically parse resource content into its component strings and phrases, map the parsed elements to the ontology and use the representative nodes to create a profile for the resources. Although a site administrator has the ability to manually create or edit any profile, an automated profile generation process would be very useful in cases where content changes frequently (such as with current or upcoming events that are continuously posted). The most basic method is to take each word in the content title and description, search through the ontology tree for any occurrences of the word, and pull out any of those concept nodes which contain the word. To build on this, regular expression matching is used so that the search in the ontology tree is done using the content words as stems, so for example "bio" would return "biology" and "biological". Regular expression matching is useful for finding words which are not exact matches, so it eliminates the limitation of missing relevant words which are just an extension of the query text.

3.2.1. Handling Over-representation and Word Sense Disambiguation

The problem with the described approach is that given the size and scope of the ontology tree, many less relevant nodes will be included in a profile. This is due to the nature of languages where one word can take on different meanings or be used in various contexts, called word sense disambiguation. In natural language, ambiguity occurs at all levels and pervades normal language use; humans have to constantly and subconsciously disambiguate in readings and conversation. To solve this problem in profile creation we assign an importance value between 0 and 1 to each concept node which is pulled for a given profile, and the higher the importance value the more related the concept node is to the content. The importance value for a concept node in a profile is calculated by averaging the similarity values of the concept node with all the other concept nodes in the profile. The notion of a similarity value is introduced in Section 3.3, and the exact formula for calculating an importance value for a node is shown in formula 2 in Section 3.3.1. In this manner concept nodes which are grouped together in sections of the ontology tree will receive higher weighting; rationally this makes sense because this provides higher weighting towards the main topic of the subject and takes weighting away from random matches which might occur. This algorithm would work in some cases where a word like "membrane" is mostly used in natural biology but might appear in the physics section of the ontology tree under diffusion processes. A cell biology class should pull all the biology concept nodes which contain "membrane" in addition to the concept node which pertains to physics diffusion, but then the greater number of biology nodes would help to give themselves a higher weighting. In some cases the algorithm will not work however; an example would be a class which taught programming structure; it would return many unrelated nodes from all over the tree like "brain structure" and "cell structure". Over-representation can be solved using a variety of methods which take into consideration word context.

[17] Deerwester S, Dumais S, Furnas G, Landauer T, Harshman, R. "Indexing by Latent Semantic Analysis." (U of Chicago, 1990.)

3.3. Relevance Determination between Profiles

Given two profiles (*c.f. Section 1.5*), whether they belong to users or content, Redbook uses a set of algorithms to calculate the similarity between the two profiles. The relevance value ranges from zero, representing no relation, to one, which represents identical preferences; this range was chosen to reflect the probability that there is a match between two profiles. In the next section 3.3.1, the complete set of algorithms for calculating relevance determination between profiles is shown and represented formulaically. The following sections 3.3.2 to 3.3.4 then explain the theories and methods used in calculating a relevance value between two profiles; the sections are split so that the algorithms are described starting from the lowest level dealing with individual concept nodes to the highest level where two profiles are compared.

3.3.1. Formulaic Representations for Relevance Determination

Using an ontology and two profiles created from the ontology, the relevance value between the two profiles is calculated using the following scheme:

Given a taxonomical hierarchy of concepts $\mathbf{\Omega}$ which consists of a set of nodes $\eta=\{n_1, \ldots, n_k\}$. A profile is a collection of nodes from the concept hierarchy. Thus the set of possible profiles is the powerset of η.

Profiles: $\boldsymbol{P} = \wp(\eta)$

where $\wp(.)$ represents the power set.

Given two profiles P_1 and $P_2 \in \boldsymbol{P}$, our goal is to find a relevance metric $R(P_1, P_2)$ for these 2 profiles. Its derivation is given below.

Relevance: R: $\boldsymbol{P} \times \boldsymbol{P} \rightarrow (0, 1]$

$P_1 = \{ n_{11}, n_{12}, \ldots n_{1k} \}$
$P_2 = \{ n_{21}, n_{22}, \ldots n_{2m} \}$

The Least Common Ancestral Node (LCAN) for two nodes is defined as the common ancestral node with the least sum distance to the two nodes (discussed in Section 3.3.2).

Similarity between 2 nodes is defined as:

$$s(n_i,n_j)=\frac{\alpha}{TD-1}LD(n_i,n_j)+\beta \quad \text{(Formula 1)}$$

where (α , β) are chosen in our implementation to be (0.9, 0.1). These values are chosen to ensure that the similarity values have a linear relationship ranging from a value of 0.1 at the top level to a value of 1.0 at the bottom level of the tree. More discussion about similarity values is in Section 3.3.3.
In formula 1, *TD* is the depth of the concept hierarchy tree. $LD(n_i, n_j)$ is defined as the level depth between any 2 arbitrary nodes in a tree:

LD: $\eta \times \eta \rightarrow (0, 1]$

The specific functional form is defined with regard to the Least Common Ancestral Node of the 2 nodes n_i and n_j so that $LD(n_i, n_j)$ = depth of their LCAN node from the root. TD and LD are illustrated more clearly in Section 3.3.3.

Importance of a node in a given profile is defined as:

Importance: $I(n_{ij}) : \eta \rightarrow (0, 1]$

The importance of a node gives a measure of how much the node pertains to the topic of the content which the profile represents.

In our implementation, we use the following functional form of *I*(.):

$$I(n_{1j})=\frac{1}{k}\sum_{n\in\eta_1}s(n_{1j},n) \quad \text{(Formula 2)}$$

We define an additional notion: the relevance of a node with respect to another profile P_2:

Relevance of a node: $R(n_{ij}, P_2) : \eta \times P \rightarrow (0, 1]$

The relevance of a node to a profile denotes how similar the node is to the set of nodes which represents the profile.
In our implementation, we use the following functional form of *R*(.) with respect to profile P_2 with concept nodes η_2:

$$R_{P_2}(n_1)=\frac{1}{m}\sum_{n_j\in P_2}I(n_j)s(n_1,n_j) \quad \text{(Formula 3)}$$

The formula to calculate the relevance of a node to a profile uses Formula 1 and Formula 2 to calculate similarity and importance for the nodes.
Finally, the relevance metric between the two profiles can be defined as:

$$R(P_1, P_2) = \frac{1}{k} \sum_{n_1 \in P_1} I(n_1) R_{P_2}(n_1) \quad \text{(Formula 4)}$$

The formula for relevance between two profiles depends on Formula 2 and Formula 3 to calculate the importance of nodes in P_1 and the relevance of individual nodes in P_1 to nodes in P_2. Section 3.3.4 presents further discussion of the methodology behind finding a relevance value between two profiles.

3.3.2. Least Common Ancestral Node

The first step in quantifying the relevance between two profiles is to establish an instrument for measuring the similarity between two concept nodes in the same ontology tree. The second step is to translate this value into a weight that can be applied to establish a relevance ranking. The algorithm used in Redbook to measure similarity between nodes is based on the depth of the LCAN (Least Common Ancestral Node) of two concept nodes. The similarity between the two nodes is directly proportional to the depth of the nodes' LCAN, given the structure and properties of an ontology tree. The specificity of the concepts increases as you go deeper into a tree, therefore nodes with a LCAN at the bottom of the tree should have more similarity than nodes with a LCAN at the top of the tree. Some limitations of using a LCAN algorithm is that it assumes that the ontology tree is not a poly-tree and can have no cycles. If a node had more than one parent, it would be unclear which path to follow when trying to find the LCAN, and levels of the tree would not be clearly established because a node could have more than one depth.

One argument made is that the similarity between two concept nodes should somehow be related to the distance between them, with one jump referring to the distance from a node to its parent or any of its children.

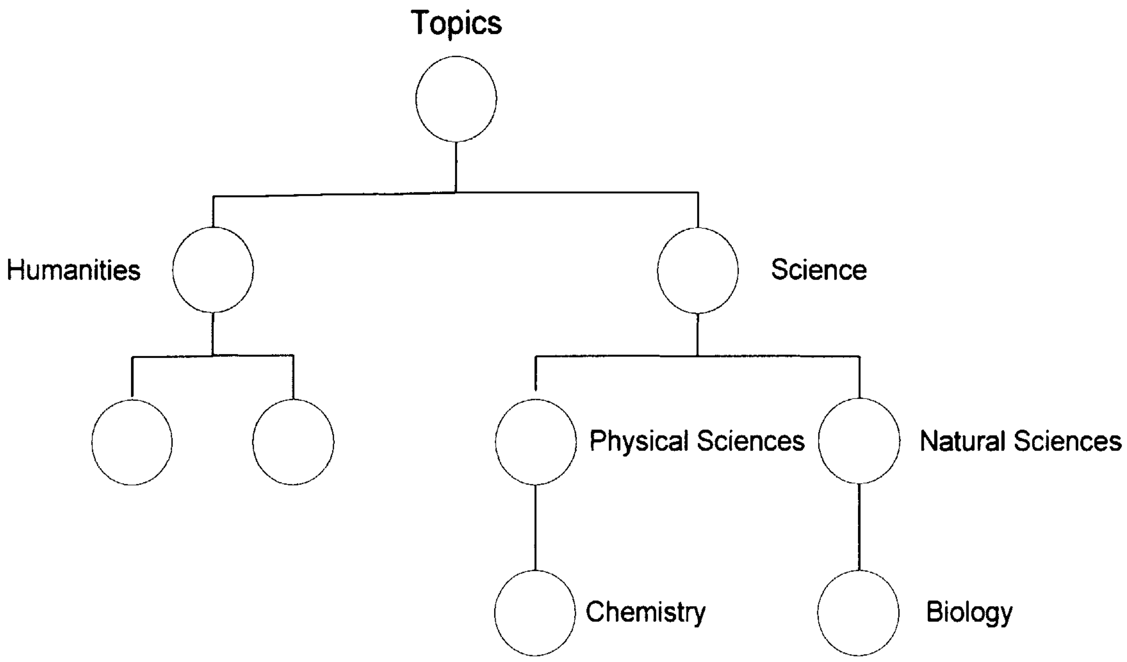


Fig. 6- Research Topics Ontology

In the figure above, nodes *Humanities* and *Science* are a distance of two jumps from each other, while nodes *Chemistry* and *Biology* have a distance of four jumps. The misconception is that since *Chemistry* and *Biology* are at a greater distance then they have less similarity, but indeed they are both types of node *Science,* so they must have more similarity between the two of them than *Humanities* has to *Science*. An algorithm based on the location of the LCAN has an advantage over a pure distance metric because it takes into account the fact that nodes at the bottom of the tree are much more similar than nodes at the top of the tree.

3.3.3. Similarity Calculation

The range of similarity is still from zero to one, so the limit cases are addressed first. Two nodes which have the ontology root node as their LCAN have a similarity of 0.1, since no two nodes in an ontology tree will ever be completely unrelated. Two identical nodes will have a similarity weighting of 1.0, but only if they are a node at the very bottom of the tree at maximum depth. For all other levels in between a linear relationship is used to assign a fraction between 0.1 and 1.0, which represents a range of 0.9 from root to maximum depth.

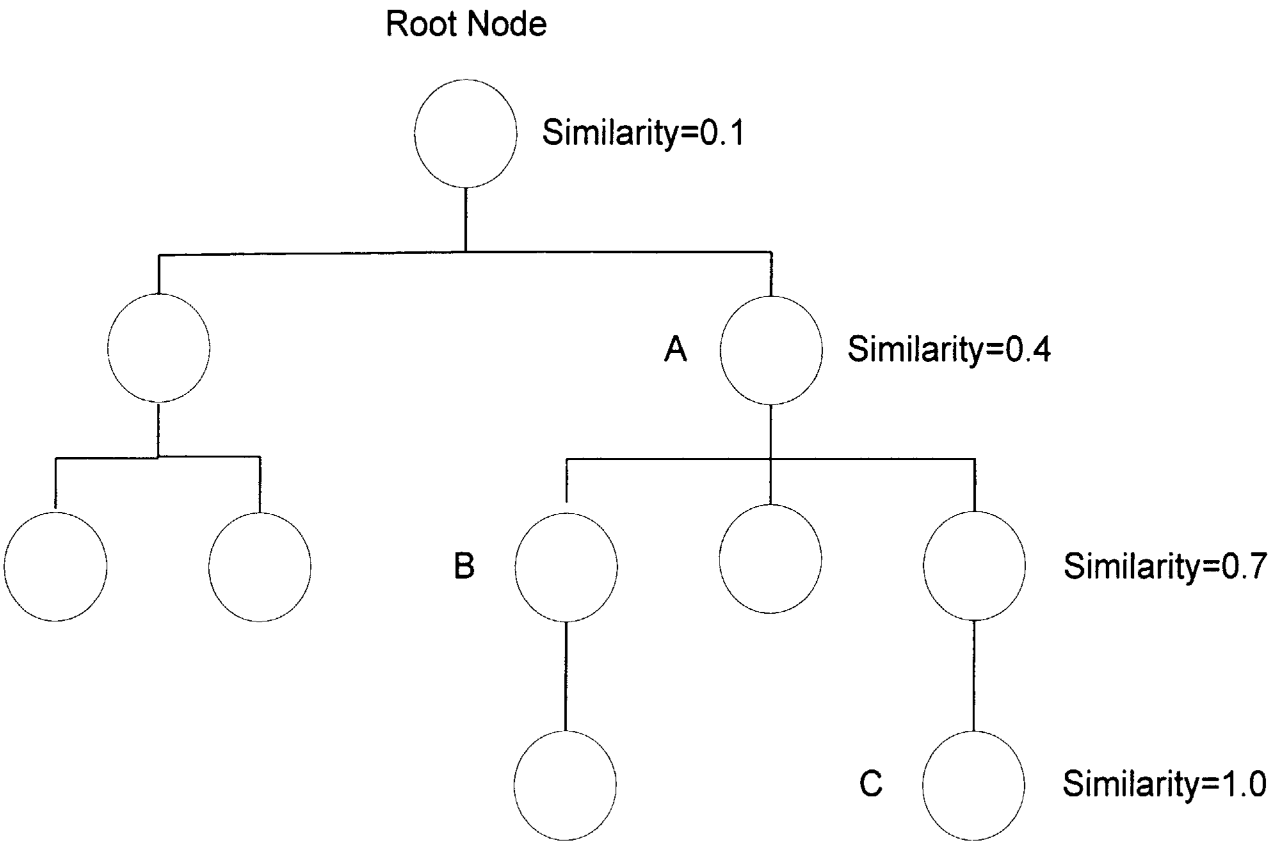


Fig. 7- Ontology Tree Similarities- At each level in the tree there is a calculated similarity value. It is the value which is the similarity between two nodes if their LCAN resides on that level.

The calculated similarity between two concept nodes is the fraction assigned to the level that their LCAN resides on, as shown in Figure 5. Based on a linear relationship, the formula used to assign values to levels in the ontology is given in Section 3.3.1 in Formula 1. Formula 1 uses the parameters TD and LD; in Figure 5, the tree depth or TD value is 4, and the level depth or LD value ranges from two to four producing similarity values of 0.4, 0.7, and 1.0 as the maximum. The similarity between nodes B and C is 0.4 since their LCAN is node A which is on a level assigned a similarity value of 0.4.

Two identical nodes which reside on a level other than the maximum depth of the tree still will not have a similarity of 1.0; rather they will follow the described algorithm and have a similarity value equal to the level they reside on. Given a tree, it is assumed that maximum information is described only at the very bottom of the tree, since those nodes are the most specific. Even if two nodes are identical at a level other than the bottom of the tree, the information provided at that level is not specific enough to have a similarity of 1.0. So in Figure 5, the similarity of node A with itself is equal to the similarity of node B and C.

3.3.4. Relevance Calculation

To find the relevance value between two profiles, which are essentially sets of concept nodes, one set of nodes is designated as the observer profile and the other set is designated as the reference profile. Generally the observer profile will belong to the user who is interacting with the Redbook environment, and the reference profile is derived from either another user's profile or the content of a project or event whose description has been parsed and mapped to the same ontology.

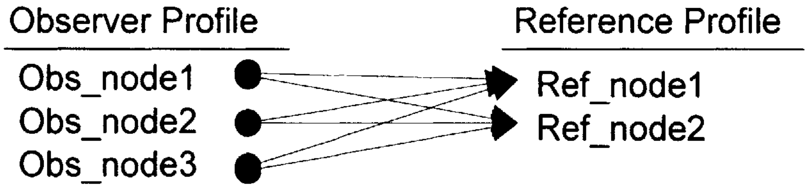


Fig. 8- Observer/Reference Profile Comparison- all nodes in the observer profile are compared to all nodes in the reference profile when determining the relevance between profiles.

A concept node in the observer profile is compared to every node in the reference profile, and the similarity values from all of these comparisons are averaged out using the reference profile's nodes' importance values as a weighting factor (Formula 3 in Section 3.3.1). These values are obtained for all observer profile nodes and then averaged across using the nodes' importance values as a weighting factor to obtain the overall relevance (Formula 4 in Section 3.3.1). Since the base similarity values between concept nodes range from 0.1 to 1.0, the relevance value between two profiles must also lie in that range.

Figure 8 shows each node in the observer profile being compared to all other nodes in the reference profile. Although currently the algorithm averages all the similarity values obtained from these comparisons, an alternative method is to just use the highest similarity value from the comparisons and discard the rest of the values. The major advantage that the current algorithm has can be illustrated using an example. Suppose the observer is interested in computer science, and the reference profile contains one computer science node and ten nodes dealing with golf and baking. Using the majority value algorithm, there would be a very high similarity value obtained since all the values from the golf and baking nodes are discarded; however if the reference user values all of his nodes equally, then the similarity between the two profiles is much lower. Averaging out all the similarity values seems to be the best way to take into account a user's full range of interests.

CHAPTER 4

PERSONALIZED RATINGS SYSTEM

The ultimate goal of my project was to create a personalized ratings system for the Redbook site; in order to accomplish this goal the architecture described in Chapter 3 had to be implemented. The figure below shows the triangular relation between a viewer, a rater, and the content, and the connections in between show how the implementation features in Chapter 3 fit into the scheme of a personalized ratings system.

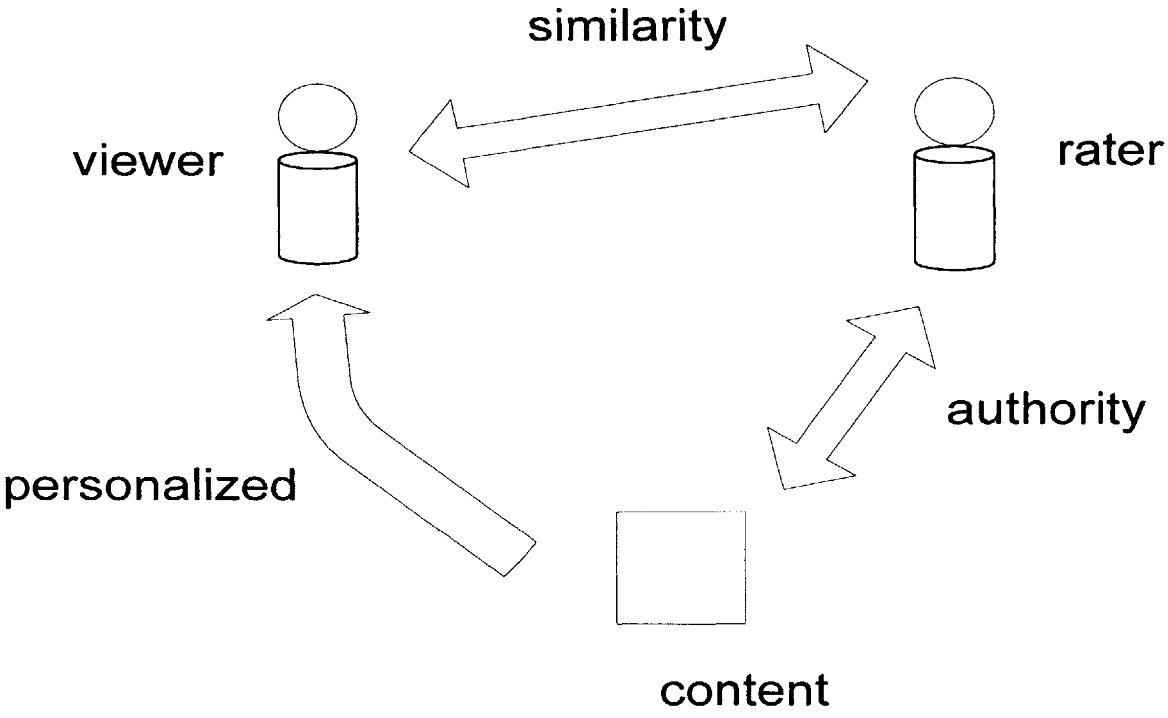


Fig. 9 Triangle relationship- A viewer looking at content on a site sees a personalized rating based on how similar he is to the raters and how much authority the raters have pertaining to the content.

The ontology-assisted profile creation for users establishes profiles between the viewer and the other raters, while automated profile creation for content can generate profiles for the content being rated. The relevance determination algorithms then tie these profiles together by calculating a "similarity" weighting between viewer and raters and an "authority" weighting between raters and content. In the figure the term "similarity" is not used in the exact sense as in the previous chapter where a similarity value was assigned to each level, rather "similarity" here refers to how similar two users are. The more similar a rater is to the viewer based on their interests specified in their profiles, the higher the weighting assigned to that rater and the higher the "similarity" between the two users. The term "authority" is used to signify the relevance calculation between the rater and content; the higher the relevance calculation between their profiles, the more "authority" the rater is said to have the matter. Both "similarity" and "authority" are then used to weight all the ratings so that the overall rating which is presented to the viewer for the content is personalized to his interests.

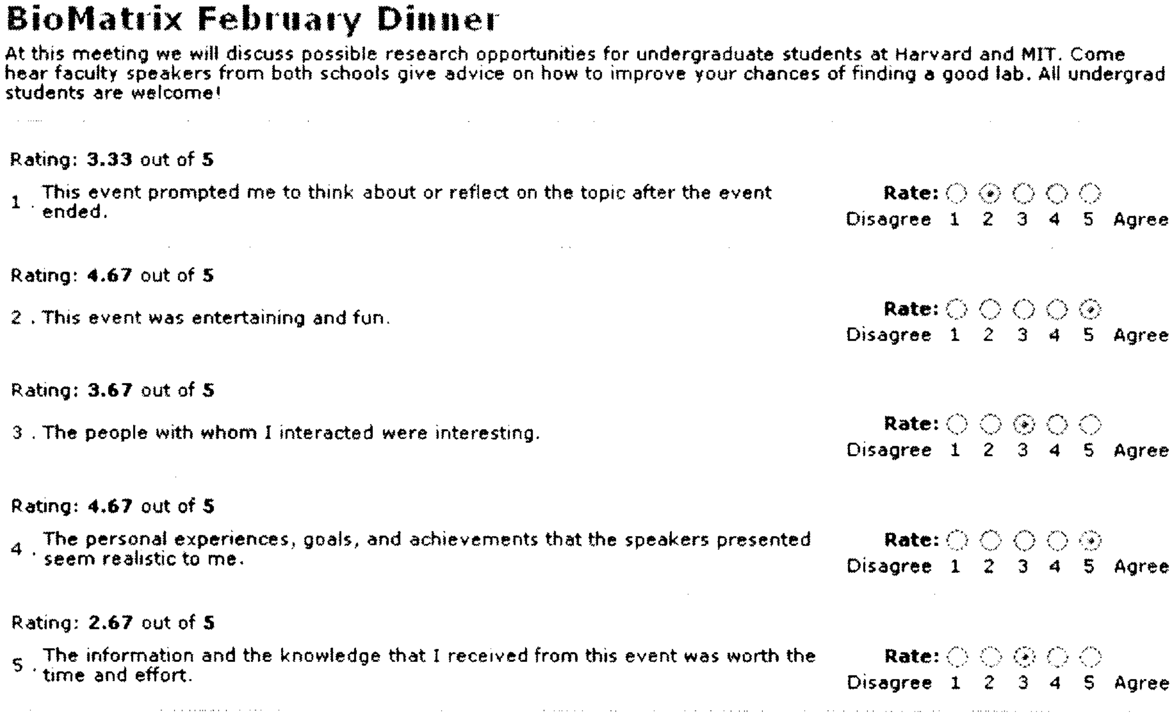
Events in Redbook are not given an overall rating by users, rather each event has statements pertaining to the quality of certain event "aspects" like organization, time value, and presentation, and users rate these statements. The method presented above where "similarity" and "authority" are used for relevance determination between profiles is used to weight the ratings for each event statement. To obtain an overall rating for each event, the event's weighted statement ratings are combined in a manner which takes into account how much weight the viewer places on aspects of an event. Section 4.1.2 will discuss event aspects in more detail.

4.1. Events Rating Implementation

The following sections describe how an event is rated and how this rating data is used to display weighted and overall rating values to the user.

4.1.1. Rating Interface

Currently personalized ratings are implemented in the events section of Redbook. Users do not give an overall rating to an event; rather there are a series of statements about the event which users are supposed to rate. This allows users to judge the quality of an event along different lines, like the entertainment value, the effectiveness of the speakers, and how interesting the topics were. These qualities of an event are referred to as "aspects." When rating an event, a series of statements about the event are presented, and the rating scale is from one to five. The following interface is presented to the user when they click to rate an individual event from the event home page which lists all the events.



BioMatrix February Dinner

At this meeting we will discuss possible research opportunities for undergraduate students at Harvard and MIT. Come hear faculty speakers from both schools give advice on how to improve your chances of finding a good lab. All undergrad students are welcome!

Rating: **3.33** out of **5**

1 . This event prompted me to think about or reflect on the topic after the event ended.

Rate: ○ ● ○ ○ ○
Disagree 1 2 3 4 5 Agree

Rating: **4.67** out of **5**

2 . This event was entertaining and fun.

Rate: ○ ○ ○ ○ ●
Disagree 1 2 3 4 5 Agree

Rating: **3.67** out of **5**

3 . The people with whom I interacted were interesting.

Rate: ○ ○ ● ○ ○
Disagree 1 2 3 4 5 Agree

Rating: **4.67** out of **5**

4 . The personal experiences, goals, and achievements that the speakers presented seem realistic to me.

Rate: ○ ○ ○ ○ ●
Disagree 1 2 3 4 5 Agree

Rating: **2.67** out of **5**

5 . The information and the knowledge that I received from this event was worth the time and effort.

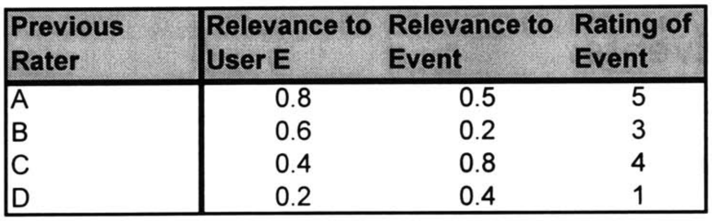
Rate: ○ ○ ● ○ ○
Disagree 1 2 3 4 5 Agree

Fig. 10- Event Rating- A sample set of statements to be rated for the event "BioMatrix February Dinner." Each statement displays the weighted rating value from all previous raters and a set of radio

buttons which gives the current rater an option of rating on a one to five scale.

A user rates an event along each of these statements, and his ratings are stored. For each statement, the current rating that is presented to the user is a weighted average from all the previous ratings from other raters. The weighting assigned to a rating from a specific rater is based on the previous discussion on "similarity" and "authority" between users and content. The scenario below demonstrates how exactly the weighting works:

User E rates an event, and one of the statements contains the following rating data for past raters A,B,C, and D:



Previous Rater	Relevance to User E	Relevance to Event	Rating of Event
A	0.8	0.5	5
B	0.6	0.2	3
C	0.4	0.8	4
D	0.2	0.4	1

The relevance values between User E and the other users and event are calculated using the algorithm described in Section 3.3.3 between two profiles. To clarify, all of an event's statements will automatically inherit the event's profile, so in calculating the weighted rating for a statement the event and user E's profiles are used.

The rater's final weight to be used in the calculation is found by multiplying the values of "Relevance to User E" and "Relevance to Event":

$$\textit{(Final Weight)} = \textit{(Relevance to User E)} * \textit{(Relevance to Event)}$$

The weighted rating is then equal to a weighted average of all the ratings based on the final weight calculated from each rater:

$$\textit{(Weighted Rating)} = \frac{\underline{[(.8)(.5)*5 + (.6)(.2)*3 + (.4)(.8)*4 + (.2)(.4)*1]}}{(.8)(.5) + (.6)(.2) + (.4)(.8) + (.2)(.4)}$$

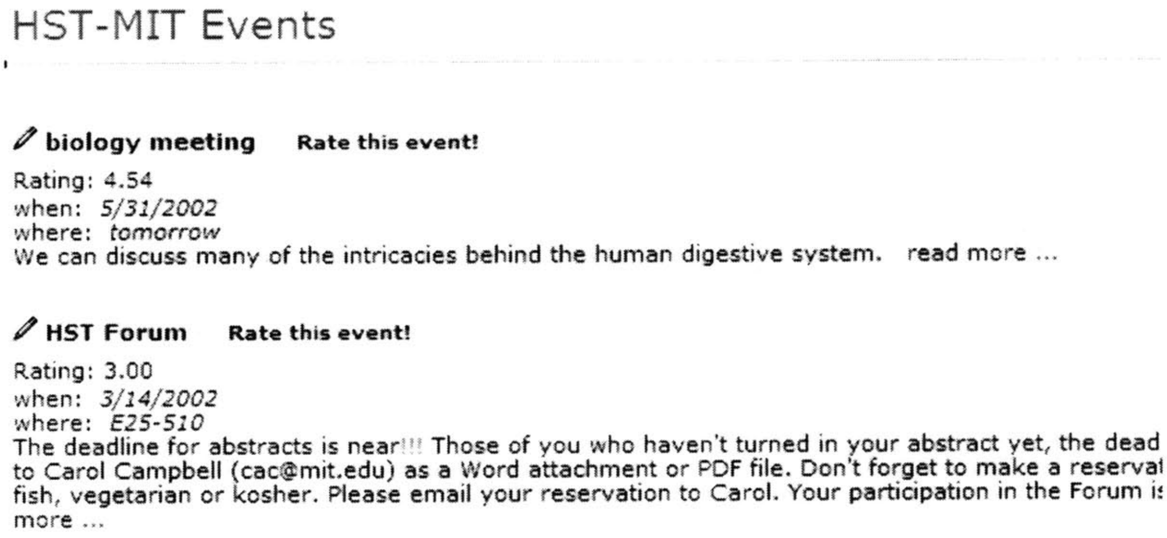
$$= 4.04$$

So the final rating that User E sees is 4.04, compared to a value of 3.25 if the ratings were just averaged without weights. Just from looking at the data table, it can be seen that the weighted rating is higher because raters A and C had higher relevance values and they rated the event 5 and 4 respectively.

The ratings displayed to the viewer for each statement is calculated using the above formula. Obtaining an overall rating for the event based on these statement ratings is detailed in the next section.

4.1.2. Event Aspects

When a user goes to the events home page, he sees a listing of events with their titles, descriptions, times and locations, and their overall ratings.



HST-MIT Events

biology meeting **Rate this event!**

Rating: 4.54
when: *5/31/2002*
where: *tomorrow*
We can discuss many of the intricacies behind the human digestive system. read more ...

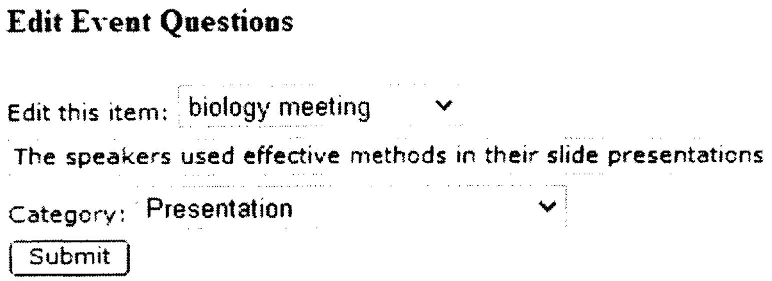
HST Forum **Rate this event!**

Rating: 3.00
when: *3/14/2002*
where: *E25-510*
The deadline for abstracts is near!!! Those of you who haven't turned in your abstract yet, the dead
to Carol Campbell (cac@mit.edu) as a Word attachment or PDF file. Don't forget to make a reservat
fish, vegetarian or kosher. Please email your reservation to Carol. Your participation in the Forum is
more ...

Fig. 11- Events Home Page- Displays a listing of events and their overall ratings, ordered by relevance to the viewer.

To calculate the overall ratings for each event displayed on the events home page, the event's statement ratings are averaged; however there is another type of weighting used in this average which doesn't have to do with relevance, and it depends on event aspects and their importance to the user which is set in the user's research profile.

As discussed before, each event has a series of statements to be rated which address different aspects of the event; these statements are chosen or created by a site administrator. An "aspect" of an event is one of various categories which reflect the characteristics an event can have, for example "time value" or "organization." During the statement creation, the administrator chooses an aspect of the event to relate the statement to; in the figure below the administrator submits a statement for the event and picks a category, or "aspect" to relate the statement to.



Edit Event Questions

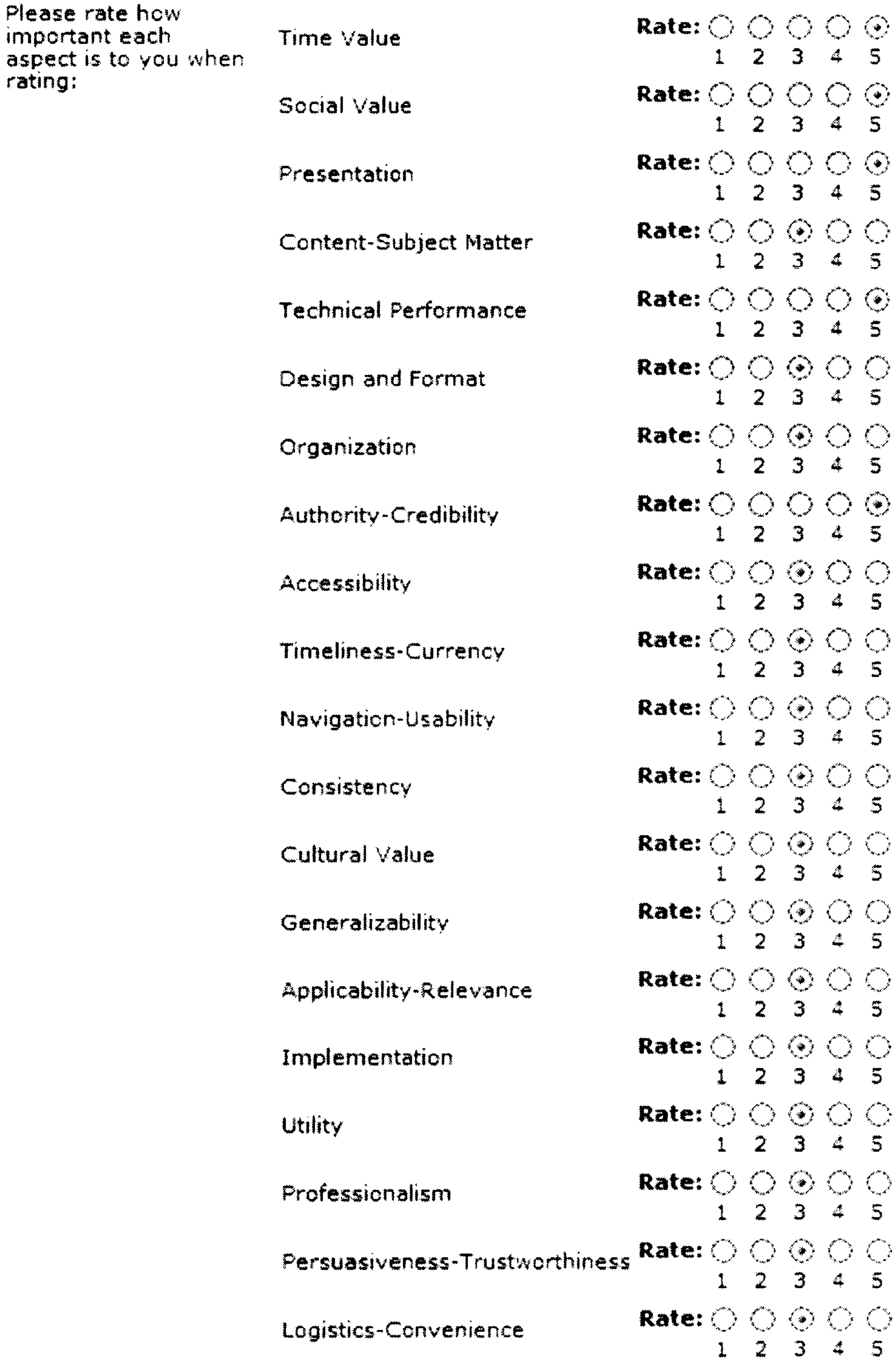
Edit this item: biology meeting

The speakers used effective methods in their slide presentations

Category: Presentation

Fig. 12- Edit Event Questions- The administrator can type in statements to add to the "biology meeting" event. Each of these statements is associated with a given category or "aspect".

Since the statements reflect different aspects of an event, there needs to be a way to quantify how important a user considers each aspect so that an overall personalized rating can be created for each user. Redbook provides a way for users to specify how important they consider the various aspects of an event. The interface below which displays all the aspects on Redbook is included in the user's research profile so that the user can edit the importance value he attributes to each aspect; the range is from one to five with a default value of three.



Please rate how important each aspect is to you when rating:

Time Value	**Rate:**	○ 1	○ 2	○ 3	○ 4	● 5
Social Value	**Rate:**	○ 1	○ 2	○ 3	○ 4	● 5
Presentation	**Rate:**	○ 1	○ 2	○ 3	○ 4	● 5
Content-Subject Matter	**Rate:**	○ 1	○ 2	● 3	○ 4	○ 5
Technical Performance	**Rate:**	○ 1	○ 2	○ 3	○ 4	● 5
Design and Format	**Rate:**	○ 1	○ 2	● 3	○ 4	○ 5
Organization	**Rate:**	○ 1	○ 2	● 3	○ 4	○ 5
Authority-Credibility	**Rate:**	○ 1	○ 2	○ 3	○ 4	● 5
Accessibility	**Rate:**	○ 1	○ 2	● 3	○ 4	○ 5
Timeliness-Currency	**Rate:**	○ 1	○ 2	● 3	○ 4	○ 5
Navigation-Usability	**Rate:**	○ 1	○ 2	● 3	○ 4	○ 5
Consistency	**Rate:**	○ 1	○ 2	● 3	○ 4	○ 5
Cultural Value	**Rate:**	○ 1	○ 2	● 3	○ 4	○ 5
Generalizability	**Rate:**	○ 1	○ 2	● 3	○ 4	○ 5
Applicability-Relevance	**Rate:**	○ 1	○ 2	● 3	○ 4	○ 5
Implementation	**Rate:**	○ 1	○ 2	● 3	○ 4	○ 5
Utility	**Rate:**	○ 1	○ 2	● 3	○ 4	○ 5
Professionalism	**Rate:**	○ 1	○ 2	● 3	○ 4	○ 5
Persuasiveness-Trustworthiness	**Rate:**	○ 1	○ 2	● 3	○ 4	○ 5
Logistics-Convenience	**Rate:**	○ 1	○ 2	● 3	○ 4	○ 5

Fig. 13- Aspect Importance- In their research profiles users are provided with an interface to choose an importance value for each aspect of an event. The scale is from one to five with a default value of three.

To calculate overall rating for an event, the following algorithm is used:

Formula 5- Overall Rating Calculation for Events

w_i – aspect importance value for statement i, chosen by user in research profile
r_i – weighted rating value for statement i, calculated using formula in example shown in Section 4.1.1

Overall event rating = $\sum (w_i * r_i) / \sum (w_i)$ (summed over all i statements)

In words, the event's overall rating is just an average of the event's statements' ratings weighted by each statement's importance value.

Aspects and importance values are used to weight the overall rating value for an event so that the ratings can be more tailored to what a user is looking for when he attends an event. It is unrelated to the relevance determination between profiles, but it is another way to personalize ratings on Redbook.

CHAPTER 5

TESTING

Superficial testing of the Redbook relevance weighting system was done using artificial agents instead of real users. Several agent profiles were created in order to simulate the interests of typical students at MIT and Harvard. The results from the testing have not been statistically validated; however, on qualitative overview, they give a good picture of how well the site will accomplish its goals:

- Ontology-assisted search and profile creation for users
- Automated profile creation for content
- Relevance determination between profiles

The following sections describe how each of the above features is tested and includes the results and some observations; more extensive analysis of the results is covered in section 6.1. Further opportunities for testing are described in section 7.2 on future work.

5.1. Ontology-assisted Search and Profile Creation

Ontology-assistance for search and profile creation will ultimately be judged successful if it can be compared favorably to a regular search and profile creation which submits only keywords specified by the user. The testing strategy for the success of this system is to fill out two sets of profiles, one which uses ontology-assistance and one which does not, and have the system match the profiles to a listing of events in rank order of preference. The returned results and their rank ordering will then be analyzed to see how well each set of profiles performed; the hypothesis is that the ontology-assisted searches will return more matches than the controlled searches since the ontology-assistance allows a user to expand his query. Matching and ranking to the events will be done with the Microsoft full-text search feature as described in Section 3.1.3. The following sections will describe the implementation of the listing of events to be tested on, the creation of both sets of ontology-assisted and control profiles, and the results and observations from the tests.

5.1.1. Event Listing

In order to artificially create a listing of events which will provide a variety of titles and descriptions in different subjects, the MIT Online Subject Listings page was parsed with classes being inserted into the database as events. In order to obtain manageable results, only classes from some of the majors were used in the testing. The five majors chosen were courses 6, 9, 15, 7, and 5, which correspond to Electrical Engineering-Computer Science (EECS), Brain and Cognitive Sciences, Management, Biology, and Chemistry, respectively. These majors were chosen because they represent a wide range of potential research topics which might be offered at MIT or Harvard.

5.1.2. Profile Creation

Given time limitations only profiles from two majors were created for both the ontology-assisted and control sets; one profile represented a student from an electrical engineering and computer science background and the other represented a student interested in the biological sciences. To create the control profile, I entered in the query "computer" for the EECS profile and "biology" for the biology profile; the interface for profile creation is described in Section 3.1.3. Based on these queries the following two profiles were created for the control set:

EECS Profile	Biology Profile
Computer	Biology

Fig. 14- EECS and Biology Control Profiles

To create the ontology-assisted profiles, I entered in the same queries as before for the EECS and biology profiles, but with the help of the ontology I added in additional terms to the profile which might be of interest to a student of the mentioned majors:

EECS Profile	Biology Profile
EECS	Biology
Computer	Molecular Biology
Electrical	Cellular Structures

Fig. 15- EECS and Biology Ontology-assisted Profiles

The ontology-assisted EECS profile contains two additional terms "EECS" and "Electrical" which are not present in its control profile, while the biology profile contains additional terms "Molecular Biology" and "Cellular Structures".

5.1.3. Results and Observations

The terms in the profiles are used in a full text search on the classes' titles and descriptions in order to find matches and obtain a rank for each match; the ranked results for the EECS ontology-assisted profile are listed in Figure 10 with their titles, course numbers, and ranks.

Title	Rank
6.090-6.094 Special Subjects in Electrical Engineering and Computer Science	64
6.100 Electrical Engineering and Computer Science Laboratory	64

6.185-6.187 Special Laboratory Subjects in Electrical Engineering and Computer Science	64
6.042J Mathematics for Computer Science	48
6.033 Computer System Engineering	48
6.035 Computer Language Engineering	48
6.001 Structure and Interpretation of Computer Programs	48
9.611J Natural Language and the Computer Representation of Knowledge	48
5.64 Biophysical Chemistry	32
9.520 Networks for Learning: Regression and Classification	32
6.245 Multivariable Control Systems	24
6.152J Microelectronics Processing Technology	21
6.012 Microelectronic Devices and Circuits	21
15.070 Advanced Stochastic Processes	21
9.02 Brain Laboratory	16
9.33 Methods in Neural Modeling	16
9.73 Visual Cognition	16
6.021J Quantitative Physiology: Cells and Tissues	16
6.003 Signals and Systems	16
15.034 Applied Econometrics and Forecasting for Management	16
15.060 Data, Models, and Decisions (Revised Content)	16
15.063 Management Decision Support Models	16
15.066J System Optimization and Analysis for Manufacturing	16

Fig. 16- EECS Ontology-assisted Profile Full-Text Search Results

The rank value is calculated by the full-text search feature using information from the inverted index rows, like word-frequency and document length. Rank can range anywhere from 0 to 1000, but the numbers have no absolute value or meaning and are only relevant relative to the other results returned in the query. The ranked results for the biology profile are listed in Figure 11 in a similar fashion.

Title	Rank
6.004 Computation Structures	96
6.024J Molecular, Cellular, and Tissue Biomechanics	64
7.29J Cellular Neurobiology	64
7.68J Cellular and Molecular Neurobiology (New)	64
9.09J Cellular Neurobiology	64
9.16 Cellular Neurophysiology	64
9.175J Cellular and Molecular Neurobiology (New)	64
9.530 Cellular and Molecular Computation	64
5.53 Molecular Structure and Reactivity I (Revised Content)	32
7.58 Molecular Biology (New)	32
7.60 Cell Biology I	32
7.28 Molecular Biology	32

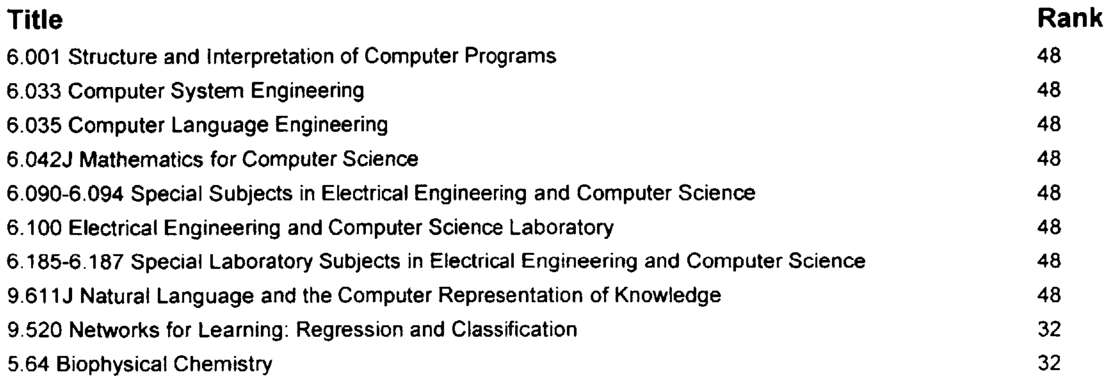
7.31 Current Topics in Mammalian Biology: Medical Implications	32
7.37J Molecular and Engineering Aspects of Biotechnology	32
7.434 Topics in Zooplankton Biology	32
7.435 Topics in Benthic Biology	32
7.436 Topics in Phytoplankton Biology	32
7.50 Method and Logic in Molecular Biology	32
7.16 Experimental Molecular Biology: Biotechnology II	32
7.22 Developmental Biology	32
5.61 Physical Chemistry	21
15.093J Optimization Methods	16
15.020 Competition in Telecommunications	16
5.07 Biological Chemistry I	16
7.05 General Biochemistry	16
9.14 Structure and Development of the Mammalian Brain	12
7.70 Regulation of Gene Expression	10
7.61 Membranes, Receptors, and Signalling	10
7.63 Immunology	10
5.062 Principles of Bioinorganic Chemistry	10
9.601J Language Acquisition I	10
9.322J Genetic Neurobiology	10
9.18 Developmental Neurobiology	10
5.43 Advanced Organic Chemistry	10
5.55 Bioorganic Chemistry (New)	10
5.08 Biological Chemistry II	10
5.111 Principles of Chemical Science	10
5.112 Principles of Chemical Science	10
5.22J Biotechnology and Engineering	10
7.UR Undergraduate Research	10
7.20 Human Physiology	10
7.13 Experimental Microbial Genetics	10
7.47 Biological Oceanography	10
7.40 Biotechnology: Engineering of Macromolecules	10
7.23 General Immunology (Revised Content and Units)	10
7.27 Principles of Human Disease	10
5.68J Kinetics of Chemical Reactions	10
5.72 Statistical Mechanics	10
5.76 Modern Topics in Physical Chemistry	10
5.79J Glycomics (New)	10
6.021J Quantitative Physiology: Cells and Tissues	8

Fig. 17- Biology Ontology-assisted Profile Full-Text Search Results

The EECS profile matching shows that the top six most relevant matches are all from the Electrical Engineering and Computer Science department, followed by a mix of classes from the science related majors, Brain and Cognitive Sciences and Chemistry.

The last results which show up are some classes from the Management department. For the biology profile, the top nine results are a mix of classes from EECS, Brain and Cognitive Sciences, Chemistry, and Biology, but after that the list is mostly dominated by classes from the Biology department. In both sets of results some of the classes were not included, thus being regarded as totally irrelevant to the user. The full-text search results seemed to miss a few appropriate classes which did not include words from the profile but still related very well. For example *15.093J Optimization Methods* deals heavily with network optimizations and theories behind computer science, but since it does not explicitly contain the word "computer" or "electrical" it was not included in the results for the EECS profile in Figure 11. Failures like this one can be amended in the future when further refinement is done to the Redbook ontology. Currently, the section of the ontology which deals with Electrical Engineering and Computer Science is not very developed and thus does not contain a full knowledge representation of that area of research.

To demonstrate the capabilities of a search facility that does not use an ontology-based profile, the same tests were run on the control sets. The control searches returned fewer results as compared to the ontology-assisted profile searches, because with fewer terms a lot of the classes simply weren't deemed relevant.



Title	Rank
6.001 Structure and Interpretation of Computer Programs	48
6.033 Computer System Engineering	48
6.035 Computer Language Engineering	48
6.042J Mathematics for Computer Science	48
6.090-6.094 Special Subjects in Electrical Engineering and Computer Science	48
6.100 Electrical Engineering and Computer Science Laboratory	48
6.185-6.187 Special Laboratory Subjects in Electrical Engineering and Computer Science	48
9.611J Natural Language and the Computer Representation of Knowledge	48
9.520 Networks for Learning: Regression and Classification	32
5.64 Biophysical Chemistry	32

6.245 Multivariable Control Systems	24
15.034 Applied Econometrics and Forecasting for Management	16
15.060 Data, Models, and Decisions (Revised Content)	16
15.063 Management Decision Support Models	16
15.066J System Optimization and Analysis for Manufacturing	16
9.02 Brain Laboratory	16
9.33 Methods in Neural Modeling	16
6.003 Signals and Systems	16
9.73 Visual Cognition	16
6.021J Quantitative Physiology: Cells and Tissues	12

Fig. 18- Control EECS Profile results

Title	Rank
7.06 Cell Biology	32
7.11 Biology Teaching	32
7.16 Experimental Molecular Biology: Biotechnology II	32
7.17 Experimental Molecular Biology: Biotechnology III	32
7.18 Topics in Experimental Biology (New)	32
7.22 Developmental Biology	32
7.28 Molecular Biology	32
7.31 Current Topics in Mammalian Biology: Medical Implications	32
7.434 Topics in Zooplankton Biology	32
7.435 Topics in Benthic Biology	32
7.436 Topics in Phytoplankton Biology	32
7.ThG Graduate Biology Thesis	32
7.50 Method and Logic in Molecular Biology	32
7.58 Molecular Biology (New)	32
7.60 Cell Biology I	32
7.340-7.349 Advanced Undergraduate Seminars	21
7.437 Topics in Molecular Biological Oceanography	16
7.439 Topics in Marine Microbiology	16
7.47 Biological Oceanography	10
7.UR Undergraduate Research	10
7.61 Membranes, Receptors, and Signaling	10
9.011 The Brain and Cognitive Sciences I	10
9.601J Language Acquisition I	10
5.062 Principles of Bioinorganic Chemistry	10
5.22J Biotechnology and Engineering	10
5.55 Bioorganic Chemistry (New)	10
5.79J Glycomics (New)	10

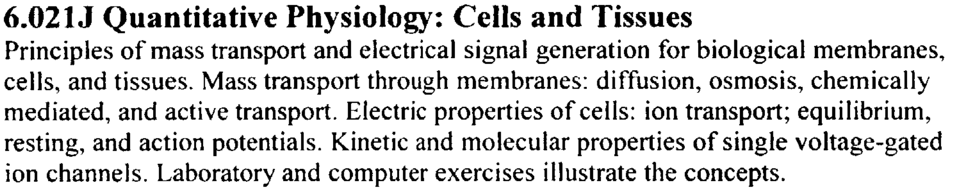
Fig. 19- Control Biology Profile results

Although the results returned were fewer, the order still seems to be very good. Noticeably the controlled biology profile results shown in Figure 19 return a series of biology classes followed by two cognitive science classes and four chemistry classes.

5.2. Automated Profile Creation

One of the goals of the system was to be able to insert content into a database and automatically generate a profile of concept nodes from the ontology given the title and description of the content. An ideal profile would be one which not only contained relevant concept nodes but excluded those concept nodes which contained keywords in the title and description but in a different context. Pulling all relevant concept nodes was accomplished by using a string search which would search in the ontology for the keyword in addition to all words which contained the keyword in it. The much more difficult task of excluding irrelevant concept nodes was tackled by calculating an importance value for each node in a profile (Formula 2 in Section 3.3.1).

To measure the system's ability to intelligently create content or resource profiles, we analyze the concept nodes which were generated for a class dealing with bio-electrical engineering. Figure 20 shows the title and description of the class *6.021J*, which encompasses both biology and electrical engineering concepts.



6.021J Quantitative Physiology: Cells and Tissues

Principles of mass transport and electrical signal generation for biological membranes, cells, and tissues. Mass transport through membranes: diffusion, osmosis, chemically mediated, and active transport. Electric properties of cells: ion transport; equilibrium, resting, and action potentials. Kinetic and molecular properties of single voltage-gated ion channels. Laboratory and computer exercises illustrate the concepts.

Fig. 20- 6.021J Title and Description

Appendix A shows the concept nodes which were generated given the title and description for *6.021J*. The concept nodes are ordered by their importance values, which are listed next to the nodes' path names. Since the automated profile creation test is more performance testing rather than experimenting with set variables and controls, there is no hypothesis, but the results will be analyzed to see if appropriate and relevant nodes were selected.

5.3. Relevance Determination between Profiles

The relevance determination testing measures the ability of the Redbook site to assign a relevance weighting between any two profiles. The basic setup is identical to the search tests in Section 5.1 where the two ontology-assisted profiles for EECS and biology are matched up against a list of classes from the MIT Online Subject Listings. Instead of using the ranking returned by Microsoft Full-Text search however, the ordering is determined by the relevance value calculated from matching profiles. The results will be analyzed in the Evaluation Chapter to see if a proper ordering was established by the relevance determination algorithms.

The results for the EECS profile are in Appendix B while Appendix C contains the results for the biology profile. The relevance rankings for the EECS results seem to be a random assortment of classes, although upon closer inspection there is a greater number of EECS classes in the top half than the bottom half, while the Management classes tend to cluster towards the bottom half. Similarly, the biology profile results show a higher tendency for biology classes to be near the top of the ranking while management classes tend to fall nearer the bottom of the ranking.

CHAPTER 6

EVALUATION

The evaluation section will first analyze the success of the three main architectural goals of my project. It will then wrap up with a discussion of the reliability of the personalized ratings system for Redbook based on how well the three features were implemented and other factors. Since the testing was preliminary and less than extensive, most of the evaluation is done using observation and intuition rather than formulaic analysis.

6.1. Architectural Goals

6.1.1. Ontology-assisted Search and Profile Creation

The ontology-assisted biology profile search returned seventy matching results, almost three times as many as the control biology profile. Allowing the user to pull out the terms "molecular biology" and "cellular structures" generated many more classes of interest which would otherwise have been hidden from just a search for the term "biology". Although the control search results seemed to be in good order with mostly biology classes followed by a few cognitive sciences and chemistry classes, the scope of the classes was very narrow and didn't cover a lot of the relevant classes in other majors. The second ranked result for the non-controlled biology profile, *6.024J Molecular, Cellular, and Tissue Biomechanics*, might very well be a research interest for a student interested in biology although the title and description do not explicitly mention the term "biology", and so it was not included in the controlled biology profile results in Figure 19. Not only did the ontology-assisted search return a broader range of classes spanning to other majors, but it also returned many more results from within the biology major.

Some failures that occurred can be seen in the ontology-assisted biology results in Figure 17. One of the specified interests in the biology profile was "cellular structures", which induced classes like *6.004 Computation Structures* and *5.53 Molecular Structure and Reactivity* to be ranked very high. The word "structure" was very common in the ontology tree and could pertain to many topics; words like "structure" in the ontology provide problems because they create a lot of noise in searches and profile creations. Further refinement of the ontology is needed in order to eliminate this type of noise in the system.

The advantages of having ontology-assistance were not as dramatic for the EECS profile, as they returned twenty-three results as opposed to twenty for the control. However, some important and relevant classes like *6.012 Microelectronic Devices and Circuits* and *6.152J Microelectronics Processing Technology* were missing from the control results, because they didn't contain the word "electrical" which had been selected when the user had ontology assistance.

The reason for the less substantial differences between ontology and control results for the EECS profile is that the word "computer" was the keyword which hooked in all the classes, and the terms "electrical" and "EECS" found from the ontology provided little additional information. On the other hand the biology profile generated many additional and various class results just from the inclusion of two terms "cellular" and "molecular biology" found from the ontology. The second case illustrates a very likely scenario on Redbook in which an undergrad doesn't have a clear picture of where his interests might lie, so he just types in the name of his major "biology." With the ontology, he can expand his search meaningfully by including related terms which might not have been to his immediate knowledge. For this reason, the ontology should prove to be a great assistance in personalizing the site for users in profile creation and project searches.

6.1.2. Automated Profile Creation

There were eighty-two generated concept nodes for the class *6.021J*, and the importance values ranged from 0.125 to 0.92. The three highest nodes with the highest importance were "Active Membrane Transport", "Electrical Properties of Membranes", and "Ion Channels"; these nodes in fact have very high relevance to the material in *6.021J*, which deals heavily with principles of mass transport through cell membranes. However, the next three nodes which follow are "Mass Fragmentography", "Singlet states", and "Difference Equation", which have little to nothing to do with the topics covered in *6.021J*. Also, the node "Membrane potentials" is fairly significant but has a 0.603 importance value, which puts it in the middle of the importance ranking because it is in the physical sciences branch of the ontology instead of the engineering or natural sciences branches.

In general the system picked up a lot of concept nodes dealing with biology and engineering, and the most appropriate concept nodes in the ontology tree for *6.021J* were ranked in the top three. The problem of intelligently finding relevant concept nodes based on the context of a class title and description is not a trivial task and should be explored further. Refinements could be made to the algorithms used for determining a concept node's importance, and the design of the ontology tree could also be improved so that the "noise" in finding relevant concept nodes is reduced. Although there are some complex algorithms which could be of use, a simple and possibly effective method is to impose a minimum relevance threshold so that only the top nodes are included.

6.1.3. Relevance Determination between Profiles

The relevance determination testing produced ranked listings which showed a slight aggregation of relevant classes closer to the top of the rankings. In general the distinction was not as clear as it was with the full-text search rankings, since the full-text search rankings produced a very definite favoring of computer classes with the computer profile and biology classes with the biology profile. However, the rankings did show some relevance in the profiles to all the classes in the subject listings, unlike the full-text search rankings which gave many classes a ranking of zero relative to the profiles.

Most of the relevance calculations seemed to be very close, with not much variation from top to bottom of a ranking list relative to a profile. This can probably be attributed to a couple factors, one being the science and engineering nature of both the ontology tree and the classes represented in the MIT Online Subject Listings. Since almost all the profiles and classes were based on science or engineering topics, it would be appropriate for all the relevance values to be very close in value. In the future if Redbook is expanded to encompass content of a non-scientific nature, the current system's algorithms would be able to support the additions.

With respect to the intelligent profile creation tests, Redbook's current algorithm system produces a lot of "noise", concept nodes which do not pertain to the content topic. The list of concept nodes for the *6.021J* class reached eighty nodes, which is probably too many to be able to accurately describe just one class. Again, thresholds could be established so that only a certain portion of top nodes are included in the calculations.

The problem of pulling two many nodes can be attributed to word sense disambiguation (WSD) as discussed in Section 3.2.1. WSD has been treated in many different ways, with varied degrees of success. Latent semantic indexing (LSI), which was briefly mentioned before, treats the unreliability of term-document association as a statistical problem[17]. A large matrix of term-document association data is used to construct a semantic space where greater relevance between document and term results in closer proximity. Singular value decomposition is then used to simplify the space by ignoring less used data and reflecting the major associative patterns in the data. Indexing would then be accomplished by position in the space, where words in a query pinpoint a location in the space and documents in the area are returned as relevant[18]. In Redbook, LSI could be set up by associating the concept nodes in the ontology tree with terms, since for our purpose the concept nodes serve as the documents to be searched through.

An improvement on LSI is Probabilistic latent semantic indexing, or PLSI. PLSI uses the fundamentals of LSI except it adds a solid statistical foundation based on the likelihood principle, and it defines a proper generative model of the data. The factor representation obtained by PLSI allows for the distinguishing between different meanings and different types of word usage.[19] As with LSI, PLSI could be set up on Redbook by constructing a semantic space using terms and the concept nodes in the ontology.

Redbook's current algorithm of weighting concept nodes to solve ambiguity in term-document association is less than ideal, since currently a description of one hundred words will contain many words which are not directly related to the topic. What renders the algorithm workable, though, is that the underlying vocabulary from the biomedical

[18] ibid

[19] Hofmann, T. "Probabilistic Latent Semantic Indexing." (UC Berkley, 1999.)

engineering domain is relatively restricted and unambiguous. Also, certain common words like “engineering” and “science” have been filtered out of the ontology; this reduces over-representation somewhat. However the issue still exists, and given more time, methods such as LSI and PLSI should be implemented on Redbook to achieve better relevance matching. More filters could also be applied so that words which exceed a certain number threshold in the ontology could be skipped over and thus suppressed.

The relevance matching system worked fairly well despite the shortcomings in the profile creation system. Given that all the profiles contained a lot of concept nodes distributed somewhat evenly throughout the ontology tree, the algorithm returned a consistent set of relevance weightings which were close in value. The average value seemed to lie at around 0.5, so when the profile creation system is improved to eliminate the noise, there will be a good distribution of values from zero to one.

6.2. Personalized Ratings System

The reliability of a personalized ratings system is dependent on a number of factors. First, users must be able to classify themselves (or self-describe their interests, expertise, etc.) using a set of concepts in the ontology tree; for this purpose Redbook provides a useful interface for users in which the ontology can be browsed to help in the profile creation process. The importance of creating profiles from an ontology as opposed to allowing users to type their interests in a textbox is that the standardized form of an ontology profile allows for a reliable system of quantifying similarities between profiles.

Second, users must be able to update their profiles when their preferences change, since interests and priorities may change quite a bit throughout an academic career. Each time a user browses the site, there should be sufficient prompts on the research and events pages for the user to update their profile. The prompts could be simple titles on the pages which simply say "Please remember to update your profile."

Finally, there must be relatively good correlation between what a user "says" and what they "do." The use of profiles to determine similarity between users is fairly unique; most sites on the Internet use ratings or purchases data to calculate user similarities. A weakness of using profiles instead of database records to determine similarity is that predicting a user's interests based on what they "say" may be less accurate than looking at what they "do." In developing our system we assume that the consistency between self-created profiles and actual interests is fairly reliable. Section 7.2 describes future work on implementing experience-based reputation to complement the profile relevance system on Redbook.

CHAPTER 7

FUTURE WORK

7.1. Testing with Real Users

The most accurate way to determine whether or not the personalized ratings assist users in their search for desired content on Redbook is to run tests with actual students and faculty in the Boston medical community. There are two main approaches which could be used in analyzing the effectiveness of the ratings system.

The first method involves setting up a situation where the users rate an event on Redbook which they have mutually attended. A week later when all the data are collected, the users are asked to revisit the site and look at two sets of ratings for the statements in an event, one which is personalized and one which is not. Without knowing which one is personalized, the user chooses which ratings he agrees with more. Given this data, the percentage of personalized ratings which received approval can be calculated to see how much better they were in conveying an accurate rating. Using this metric the approval rating should most likely be over fifty percent, and a good goal to aim for would be a seventy-five percent approval rating.

A different approach for testing the ratings system also starts with having users rate an event which they have attended. Instead of having the users judge the personalized data, the actual rating data is compared to the weighted rating values which would have been displayed to the rater. Using these two lists of rating data, the Wilcoxon Signed-Rank[20] test will be used to assess how similar the two lists are to each other. The

[20] http://faculty.vassar.edu/lowry/wilcoxon.html

hypothesis is that the personalized ratings which are to be displayed to a user should very accurately represent how the user would rate, so the two bodies of data should be similar.

7.2. Experience-based Reputation

The personalized rating system establishes a rater's "authority" based on the relevance calculation, which is fed off of what the viewer and rater generate in their profile. In other words, a user's relevance is weighted based on what he "says." Another, possibly better method of calculating a user's relevance is to observe what a user "does." Users' actions would simply be the database records of their past rating behavior in this environment, and this data can be used to calculate how similar two users are by looking at their rating trends. The more similar two users rated in the past, the higher the relevance between them is. There are currently web sites which use collaborative filtering algorithms, as they are generally called, to predict the utility of items to a particular user based on a set of user votes from the database.[21] www.Amazon.com is perhaps the most popular example of this; based on user ratings and purchases the site tailors a set of preferred items to all their users.[22] The algorithms can easily be used to calculate a correlation value between users.

[21] Breese J, Heckerman D, Kadie C. "Empirical Analysis of Predictive Algorithms for Collaborative Filtering." (Microsoft Research, 1998.)

[22] Patel N. (Lecture from 15.062 Data Mining: Algorithms and Applications, April 2002.)

References

Aggarwal C, Yu P. "On Effective Conceptual Indexing and Similarity Search in Text Data." (IBM T.J. Watson Research Center, 2001.)

Breese J, Heckerman D, Kadie C. "Empirical Analysis of Predictive Algorithms for Collaborative Filtering." (Microsoft Research, 1998.)

Deerwester S, Dumais S, Furnas G, Landauer T, Harshman, R. "Indexing by Latent Semantic Analysis." (U of Chicago, 1990.)

Faloutsos C. "Access Methods for Text." (ACM Computing Surveys 17, March 1, 1995.)

Guarino, Nicola. "Formal Ontology and Information Systems." (National Research Council, 1998.)

Hofmann, T. "Probabilistic Latent Semantic Indexing." (UC Berkley, 1999.)

Jasper R and Uschold M. "A Framework for Understanding and Classifying Ontology Applications." (Boeing Math and Computing Technology, 1999.)

Luke S, Spector L, Rager D, Hendler J. "Ontology-based Web Agents." (University of Maryland and Hampshire College, 1996.)

"Microsoft Full-Text Search Technologies: White Paper." (Microsoft Corporation, 2001.)

Patel N. (Lecture from 15.062 Data Mining: Algorithms and Applications, April 2002.)

Salton G, McGill MJ. "Introduction to Modern Information Retrieval." (McGraw Hill, 1983.)

Stubkjaer, E. "Integrating Ontologies: Assessing the Use of the Cyc Ontology for Cadastral Applications." (Aalborg University, 2000.)

Devhood. Viewed in 2001 at [http://www.devhood.com]
eBay. Viewed in 2001 at [http://www.ebay.com]
Slashdot. Viewed in 2001 at [http://slashdot.org/faq/]

http://www.ksl.stanford.edu/software/ontolingua/
http://protege.stanford.edu/
http://faculty.vassar.edu/lowry/wilcoxon.html

Appendix A- 6.021J Concept Nodes and Importance

Concept Nodes	Importance
Natural/Biology/Cell Biology/Cellular Structures/Cell Membrane/Membrane Transport/Active Membrane Transport	0.92
Natural/Biology/Cell Biology/Cellular Structures/Cell Membrane/Membrane Transport/Electrical Properties of Membranes	0.914
Natural/Biology/Cell Biology/Cellular Structures/Cell Membrane/Membrane Transport/Ion Channels	0.91
Physical/Chemistry/Chemistry, Analytical/Spectrum Analysis/Spectrum Analysis, Mass/Mass Fragmentography	0.822
Engineering/BioMedical/Bio Optics/Spectroscopy/Energy Level Diagrams/Singlet states	0.822
Mathematics/Statistics Probability/Probability Theory/Stochastic Processes/Random Walks/Difference Equation	0.819
Natural/Biology/Cell Biology/Cellular Structures/Cell Membrane/Membrane Transport	0.817
Engineering/BioMedical/Bio Optics/Light Propagation in Turbid Media/Models/Diffusion	0.814
Engineering/BioMedical/Bio Optics/Spectroscopy/Scattering/Refraction Variations	0.809
Physical/Chemistry/Chemistry, Physical/Electrochemistry/Electrophoresis/Isoelectric Focusing	0.801
Engineering/BioMedical/Bio Optics/Optical Trapping/Micromanipulation of cells	0.727
Engineering/BioMedical/BioTechnology/Cellular BioMechanics/Mechanical Properties of Tissues	0.725
Natural/Biology/Cell Biology/Cell Division/M Phase	0.724
Physical/Chemistry/Chemistry, Physical/Permeability/Osmosis	0.724
Engineering/BioMedical/BioTechnology/Separation and Analytical Techniques/Cell Isolation and Fractionation	0.722
Physical/Chemistry/Chemistry, Analytical/Crystallography/X Ray Diffraction	0.722
Physical/Chemistry/Chemistry, Physical/Crystallography/X Ray Diffraction	0.722
Engineering/BioMedical/BioTechnology/Ethics/Biological Sources	0.721
Medicine/Immunology Immunological/Immune cells/Leukocyte/Mast Cell/Allergy Allergic Reaction	0.72
Engineering/BioMedical/BioTechnology/Protein Therapeutics/Pharmacokinetics	0.72
Physical/Physics/Nuclear Physics/Magnetic Resonance Spectroscopy/Nuclear Magnetic Resonance, Biomolecular	0.72
Natural/Biology/Cell Biology/Energy Conversion/Evolution of Electron Transport Chains	0.715
Engineering/BioMedical/BioTechnology/Cellular BioMechanics/BioElectrical Events	0.714
Engineering/BioMedical/Bio Optics/Electromagnetic Radiation/Diffraction	0.714
Physical/Physics/Biophysics Biological Physics/Biomechanics/Kinetics	0.714
Engineering/BioMedical/BioTechnology/Ethics/Stem Cells	0.71
Engineering/BioMedical/Bio Optics/Electromagnetic Radiation/Refraction	0.71
Physical/Chemistry/Chemistry, Analytical/Chromatography/Micellar Electrokinetic Capillary	0.71
Engineering/BioMedical/BioTechnology/BioReactors/Mass Transfer of Oxygen Mixing	0.707
Engineering/BioMedical/BioTechnology/Cellular BioMechanics/Electrical Regulation	0.704
Physical/Chemistry/Chemistry, Analytical/Spectrum Analysis/Spectrum Analysis, Mass	0.701
Physical/Chemistry/Chemistry, Analytical/Electrophoresis/Isoelectric Focusing	0.701
Natural/Biology/Cell Biology/Cytoskeleton/Contraction	0.701
Natural/Biology/Molecular Biology/Molecular Genetics	0.624
Medical Informatics/Medical Computing Methodologies, Computational Methods/Computer Graphics/Computer-Aided Design	0.624
Physical/Physics/Biophysics Biological Physics/Phase Boundaries	0.624
Physical/Chemistry/Chemistry, Pharmaceutical/Drug Design	0.624
Physical/Physics/Psychophysics/Signal Detection (Psychology)	0.622
Physical/Physics/Mechanics/Kinetics	0.614
Medical Informatics/Medical Computing Methodologies, Computational Methods/Artificial Intelligence/Neural Networks (Computer)	0.614
Medical Informatics/Medical Computing Methodologies, Computational Methods/BioInformatics/Biological Pathwya Discovery	0.612
Physical/Chemistry/Chemistry, Analytical/Biuret Reaction	0.609

Physical/Physics/Nuclear Physics/Elementary Particle Interactions	0.609
Mathematics/Differential Equations/Partial Differential Equations/Linear Partial Differential Equations	0.609
Physical/Physics/Biophysics Biological Physics/Diffusion	0.609
Engineering/BioMedical/BioTechnology/Pharmaceutical Plant Design	0.609
Physical/Chemistry/Chemistry, Physical/Surface Properties	0.607
Natural/Biology/Cell Biology/Cell Signalling	0.605
Engineering/BioMedical/Bioinformatics - Computational Molecular Biology/Consensus - Signal Sequences	0.603
Physical/Physics/Biophysics Biological Physics/Membrane Potentials	0.603
Physical/Chemistry/Chemistry, Physical/Molecular Conformation	0.603
Physical/Chemistry/Chemistry, Physical/Molecular Structure	0.603
Physical/Chemistry/Chemistry, Physical/Molecular Weight	0.603
Physical/Chemistry/Chemistry, Analytical/Differential Thermal Analysis	0.601
Physical/Chemistry/Chemistry, Physical/Membranes, Artificial	0.599
Physical/Chemistry/Chemistry, Analytical/Fractionation	0.599
Physical/Chemistry/Chemistry, Physical/Maillard Reaction	0.599
Medicine/Immunology Immunological/Immunity/Immunity, Antibody Mediated	0.525
Engineering/BioMedical/Bioinformatics - Computational Molecular Biology	0.525
Natural/Biology/Molecular Biology	0.525
Engineering/EECS/Computer	0.525
Engineering/EECS/Electrical	0.522
Mathematics/Nonlinear Dynamics and Chaos/Phase Plane	0.522
Medicine/Medical Informatics/Medical Informatics Applications/Computer-Assisted Clinical Decision Making	0.522
Mathematics/Differential Equations/Ordinary Differential Equations	0.522
Mathematics/Differential Equations/Partial Differential Equations	0.519
Medicine/Medical Informatics/Medical Informatics Applications/Computer-Assisted Therapy	0.514
Physical/Physics/Biophysics Biological Physics	0.512
Medicine/Medical Informatics/Medical Informatics Applications/Computer-Assisted Diagnosis	0.509
Medicine/Medical Informatics/Medical Computing Methodologies, Computational Methods/Computer Simulation, Medical Simulation	0.508
Medicine/Medical Informatics/Medical Computing Methodologies, Computational Methods/Computer Graphics	0.505
Medicine/Medical Informatics/Medical Computing Methodologies, Computational Methods/Computer-Assisted Signal Processing	0.504
Mathematics/Nonlinear Dynamics and Chaos/Phase Locking	0.504
Medicine/Nutrition/Nutritional Requirements	0.422
Medicine/Immunology Immunological/Immune cells	0.417
Medicine/Rheumatology Rheumatological Disorders/Articular Cartilage Degeneration	0.414
Mathematics/Differential Equations	0.414
Medicine/Pathology/Clinical and Laboratory Immunology	0.407
Engineering	0.327
Medicine/Connective Tissues Disorders	0.312
Science and Engineering	0.227
/imatch/Research Concepts	0.125

Appendix B- Relevance Determination EECS Profile results

Title	Relevance
6.199 Advanced Undergraduate Project	0.5
9.68 Affect: Biological, Psychological, and Social Aspects of ``Feelings"	0.4940062
15.034 Applied Econometrics and Forecasting for Management	0.4892874
7.491 Research in Biological Oceanography	0.4889753
6.100 Electrical Engineering and Computer Science Laboratory	0.4871396
9.641 Introduction to Neural Networks	0.4869807
15.071 Decision Techniques for Managers	0.4866682
15.062 Data Mining: Algorithms and Applications	0.4866192
9.364 Research in Cognitive Architectures	0.4863512
5.50 Enzymes: Structure and Function	0.4862818
6.171 Software Engineering for Web Applications (New)	0.4861194
6.231 Dynamic Programming and Stochastic Control	0.4861017
5.46 NMR Spectroscopy and Organic Structure Determination	0.486102
6.001 Structure and Interpretation of Computer Programs	0.4859878
7.18 Topics in Experimental Biology (New)	0.4859252
5.56 Molecular Structure and Reactivity II	0.4858965
6.192 Prototyping Research Results II	0.4858754
7.75J Topics in Metabolic Biochemistry	0.4857477
6.070J Electronics Project Laboratory	0.4856664
6.012 Microelectronic Devices and Circuits	0.4855784
6.011 Introduction to Communication, Control, and Signal Processing	0.4854306
5.78 Practical Macromolecular Crystallography (New)	0.4852358
6.151 Semiconductor Devices Project Laboratory	0.4851637
7.23 General Immunology (Revised Content and Units)	0.4850837
5.76 Modern Topics in Physical Chemistry	0.4850689
9.530 Cellular and Molecular Computation	0.4850718
9.15 Biochemistry and Pharmacology of Synaptic Transmission	0.4849414
7.16 Experimental Molecular Biology: Biotechnology II	0.4848975
6.170 Laboratory in Software Engineering	0.4846818
7.31 Current Topics in Mammalian Biology: Medical Implications	0.484412
6.191 Prototyping Research Results I (Revised Units)	0.4843901
6.111 Introductory Digital Systems Laboratory	0.4843171
9.03 Neural Basis of Learning and Memory	0.4842931
6.061 Introduction to Electric Power Systems	0.4842897
6.121J Bioelectronics Project Laboratory	0.4842514
7.13 Experimental Microbial Genetics	0.4841565
5.80 Special Topics in Chemical Physics	0.4841422
6.045J Automata, Computability, and Complexity	0.4841152
9.011 The Brain and Cognitive Sciences I	0.4840887
7.40 Biotechnology: Engineering of Macromolecules	0.4839543
5.49 Membrane and Receptor Biochemistry	0.4839596
7.440 An Introduction to Mathematical Ecology	0.4838923
7.68J Cellular and Molecular Neurobiology (New)	0.4838972
9.175J Cellular and Molecular Neurobiology (New)	0.4838972
5.311 Introductory Chemical Experimentation	0.4838774

7.77 Nucleic Acids, Structure, Function, Evolution and Their Interactions with Proteins	0.4838515
6.182 Psychoacoustics Project Laboratory	0.4837388
6.013 Electromagnetic Fields and Energy	0.4837129
5.111 Principles of Chemical Science	0.4836571
15.023J Global Climate Change: Economics, Science, and Policy	0.4836657
9.14 Structure and Development of the Mammalian Brain	0.4836405
5.32 Intermediate Chemical Experimentation	0.4836261
5.112 Principles of Chemical Science	0.4835638
15.063 Management Decision Support Models	0.4835028
9.34J Perception, Knowledge, and Cognition	0.4833926
6.042J Mathematics for Computer Science	0.483378
15.094 Systems Optimization: Models and Computation (Revised Units)	0.4833356
15.060 Data, Models, and Decisions (Revised Content)	0.4833012
9.520 Networks for Learning: Regression and Classification	0.4832873
7.25 Biological Regulatory Mechanisms	0.4832924
6.152J Microelectronics Processing Technology	0.483273
7.59J Teaching College-Level Science (New)	0.4832551
6.161 Modern Optics Project Laboratory (Revised Units)	0.4832441
15.064J Engineering Probability and Statistics	0.4831975
9.63 Laboratory in Cognitive Science	0.4830861
15.066J System Optimization and Analysis for Manufacturing	0.4830541
15.232 The Firm and The Business Environment in Japan	0.4830368
15.077 Modern Regression and Multivariate Data Mining	0.4830275
7.17 Experimental Molecular Biology: Biotechnology III	0.4830327
6.021J Quantitative Physiology: Cells and Tissues	0.4830077
15.076 Statistical Theory and Data Analysis	0.4829959
7.61 Membranes, Receptors, and Signalling	0.4829951
5.04 Principles of Inorganic Chemistry II	0.4829783
5.63 Molecular Spectroscopy: Laser and Magnetic Resonance Techniques	0.4829758
9.373 Somatosensory and Motor Systems	0.4829541
9.29 Introduction to Computational Neuroscience (Revised Content)	0.4827786
6.041 Probabilistic Systems Analysis	0.4827819
5.74 Introductory Quantum Mechanics II	0.4827781
9.35 Sensation and Perception (Revised Content)	0.4826723
6.022J Quantitative Physiology: Organ Transport Systems	0.4826563
6.251J Introduction to Mathematical Programming	0.4825471
5.068 Physical Methods in Inorganic Chemistry	0.4824705
6.115 Microcomputer Project Laboratory	0.4824131
9.50 Research in Brain and Cognitive Sciences	0.4823398
15.010 Economic Analysis for Business Decisions (Revised Units)	0.4822686
5.062 Principles of Bioinorganic Chemistry	0.4822471
15.072 Queues: Theory and Applications	0.4821686
15.024 Applied Economics for Managers	0.4820974
9.19J Cognitive and Behavioral Genetics (Revised Content)	0.482022
15.070 Advanced Stochastic Processes	0.4820117
15.036 Modern Econometrics for Management	0.4817935
15.233 Business Organization and Environment: Latin America	0.4817669
5.73 Introductory Quantum Mechanics I	0.4816774
9.71 Functional MRI of High-Level Vision (New)	0.4814472
6.242 Advanced Linear Control Systems	0.4813924

15.215 International Dimensions of Management	0.4813601
15.020 Competition in Telecommunications	0.4813381
15.067 Competitive Decision-Making and Negotiation (Revised Content)	0.4812566
15.013 Industrial Economics for Strategic Decisions	0.4812256
7.37J Molecular and Engineering Aspects of Biotechnology	0.4808855
5.68J Kinetics of Chemical Reactions	0.4805015
15.018 Management and Policy in the International Economy	0.4803956
9.74 Foundations of Human Memory and Learning	0.4803089
15.222 Managing International Enterprises	0.4801653
9.012 The Brain and Cognitive Sciences II	0.480089
6.291 Seminar in Systems, Communications, and Control Research	0.4797288
5.301 Chemistry Laboratory Techniques	0.4795431
15.019 International Trade and Competition	0.4792749
15.249 Special Seminar in International Management	0.4791593
15.136J Principles and Practice of Drug Development	0.4791413
15.141J Economics of the Health Care Industries (Revised Content)	0.4790912
9.081 Human Memory and Learning (Revised Content)	0.4772205
9.04 Neural Basis of Vision and Audition	0.476458

Appendix C- Relevance Determination Biology Profile results

Title	Importance
9.68 Affect: Biological, Psychological, and Social Aspects of ``Feelings"	0.5518561
7.18 Topics in Experimental Biology (New)	0.5518083
7.25 Biological Regulatory Mechanisms	0.5390418
6.171 Software Engineering for Web Applications (New)	0.5382292
5.46 NMR Spectroscopy and Organic Structure Determination	0.5373735
7.75J Topics in Metabolic Biochemistry	0.5361347
6.151 Semiconductor Devices Project Laboratory	0.5358811
5.50 Enzymes: Structure and Function	0.5355273
7.13 Experimental Microbial Genetics	0.5352707
7.16 Experimental Molecular Biology: Biotechnology II	0.5348798
6.182 Psychoacoustics Project Laboratory	0.5345645
9.530 Cellular and Molecular Computation	0.534389
5.56 Molecular Structure and Reactivity II	0.5341215
6.001 Structure and Interpretation of Computer Programs	0.5336741
15.062 Data Mining: Algorithms and Applications	0.5333089
9.011 The Brain and Cognitive Sciences I	0.5332645
5.78 Practical Macromolecular Crystallography (New)	0.5330658
6.170 Laboratory in Software Engineering	0.5329702
6.042J Mathematics for Computer Science	0.5329134
6.191 Prototyping Research Results I (Revised Units)	0.5329027
7.37J Molecular and Engineering Aspects of Biotechnology	0.532639
7.23 General Immunology (Revised Content and Units)	0.5320559
6.192 Prototyping Research Results II	0.5318584
9.74 Foundations of Human Memory and Learning	0.5317979
9.641 Introduction to Neural Networks	0.5316792
9.29 Introduction to Computational Neuroscience (Revised Content)	0.5315061
6.111 Introductory Digital Systems Laboratory	0.5313595
15.071 Decision Techniques for Managers	0.5312343
9.373 Somatosensory and Motor Systems	0.5311344
6.231 Dynamic Programming and Stochastic Control	0.5309905
5.111 Principles of Chemical Science	0.530414
5.112 Principles of Chemical Science	0.5301942
9.15 Biochemistry and Pharmacology of Synaptic Transmission	0.5298806
5.49 Membrane and Receptor Biochemistry	0.5297467
6.121J Bioelectronics Project Laboratory	0.529756
15.064J Engineering Probability and Statistics	0.5288286
7.31 Current Topics in Mammalian Biology: Medical Implications	0.5287827
15.215 International Dimensions of Management	0.5286409
9.63 Laboratory in Cognitive Science	0.5286077
9.364 Research in Cognitive Architectures	0.5285668
7.68J Cellular and Molecular Neurobiology (New)	0.5285361
9.175J Cellular and Molecular Neurobiology (New)	0.5285361
7.17 Experimental Molecular Biology: Biotechnology III	0.5283546
15.024 Applied Economics for Managers	0.5282452
9.03 Neural Basis of Learning and Memory	0.5279447

5.32 Intermediate Chemical Experimentation	0.5279036
7.59J Teaching College-Level Science (New)	0.5278039
9.520 Networks for Learning: Regression and Classification	0.5277103
15.094 Systems Optimization: Models and Computation (Revised Units)	0.5276915
5.311 Introductory Chemical Experimentation	0.5276865
6.012 Microelectronic Devices and Circuits	0.5275231
9.14 Structure and Development of the Mammalian Brain	0.5274253
15.076 Statistical Theory and Data Analysis	0.5273069
7.40 Biotechnology: Engineering of Macromolecules	0.5272887
6.022J Quantitative Physiology: Organ Transport Systems	0.5271664
5.301 Chemistry Laboratory Techniques	0.5270994
6.251J Introduction to Mathematical Programming	0.5271022
6.041 Probabilistic Systems Analysis	0.5270497
9.012 The Brain and Cognitive Sciences II	0.5270432
15.010 Economic Analysis for Business Decisions (Revised Units)	0.5270159
15.019 International Trade and Competition	0.5270022
5.76 Modern Topics in Physical Chemistry	0.526965
9.71 Functional MRI of High-Level Vision (New)	0.5268145
9.081 Human Memory and Learning (Revised Content)	0.526669
5.80 Special Topics in Chemical Physics	0.5264656
7.61 Membranes, Receptors, and Signalling	0.5264496
15.060 Data, Models, and Decisions (Revised Content)	0.52645
15.077 Modern Regression and Multivariate Data Mining	0.5264191
15.072 Queues: Theory and Applications	0.5263572
7.440 An Introduction to Mathematical Ecology	0.5263005
7.491 Research in Biological Oceanography	0.526297
15.063 Management Decision Support Models	0.5262561
9.34J Perception, Knowledge, and Cognition	0.5261726
5.63 Molecular Spectroscopy: Laser and Magnetic Resonance Techniques	0.5260707
15.066J System Optimization and Analysis for Manufacturing	0.5258326
5.068 Physical Methods in Inorganic Chemistry	0.5256063
6.045J Automata, Computability, and Complexity	0.5254781
6.021J Quantitative Physiology: Cells and Tissues	0.5254595
7.77 Nucleic Acids, Structure, Function, Evolution and Their Interactions with Proteins	0.5253729
9.50 Research in Brain and Cognitive Sciences	0.5252609
15.023J Global Climate Change: Economics, Science, and Policy	0.5251691
6.100 Electrical Engineering and Computer Science Laboratory	0.524857
15.036 Modern Econometrics for Management	0.5248382
15.222 Managing International Enterprises	0.5247934
15.013 Industrial Economics for Strategic Decisions	0.5247506
6.115 Microcomputer Project Laboratory	0.5244165
5.04 Principles of Inorganic Chemistry II	0.5241601
15.034 Applied Econometrics and Forecasting for Management	0.5241331
6.070J Electronics Project Laboratory	0.5240204
15.018 Management and Policy in the International Economy	0.5239897
5.73 Introductory Quantum Mechanics I	0.5239238
6.152J Microelectronics Processing Technology	0.5233426
15.070 Advanced Stochastic Processes	0.5232671
6.161 Modern Optics Project Laboratory (Revised Units)	0.5229697
9.35 Sensation and Perception (Revised Content)	0.5219766

9.19J Cognitive and Behavioral Genetics (Revised Content)	0.5218078
5.74 Introductory Quantum Mechanics II	0.5217793
5.062 Principles of Bioinorganic Chemistry	0.5212016
15.020 Competition in Telecommunications	0.5198097
15.249 Special Seminar in International Management	0.5194018
6.011 Introduction to Communication, Control, and Signal Processing	0.5192618
6.242 Advanced Linear Control Systems	0.5189846
5.68J Kinetics of Chemical Reactions	0.517672
15.232 The Firm and The Business Environment in Japan	0.5173191
15.136J Principles and Practice of Drug Development	0.5167343
6.061 Introduction to Electric Power Systems	0.5166183
15.067 Competitive Decision-Making and Negotiation (Revised Content)	0.5161491
15.233 Business Organization and Environment: Latin America	0.5159774
6.013 Electromagnetic Fields and Energy	0.515672
15.141J Economics of the Health Care Industries (Revised Content)	0.5153263
9.04 Neural Basis of Vision and Audition	0.5151397
6.291 Seminar in Systems, Communications, and Control Research	0.5096828
6.199 Advanced Undergraduate Project	0.5