Topic Segmentation of Message Hierarchies for Indexing and Navigation Support [*]

Jong Wook Kim
Comp. Sci. and Eng. Dept.
Arizona State University
Tempe, AZ 85287
jong@asu.edu

K. Selçuk Candan
Comp. Sci. and Eng. Dept.
Arizona State University
Tempe, AZ 85287
candan@asu.edu

Mehmet E. Dönderler
Comp. Sci. and Eng. Dept.
Arizona State University
Tempe, AZ 85287
mehmet.donderler@asu.edu

ABSTRACT

Message hierarchies in web discussion boards grow with new postings. Threads of messages evolve as new postings focus within or diverge from the original themes of the threads. Thus, just by investigating the subject headings or contents of earlier postings in a message thread, one may not be able to guess the contents of the later postings. The resulting navigation problem is further compounded for blind users who need the help of a screen reader program that can provide only a *linear* representation of the content. We see that, in order to overcome the navigation obstacle for blind as well as sighted users, it is essential to develop techniques that help identify how the content of a discussion board grows through generalizations and specializations of topics. This knowledge can be used in segmenting the content in coherent units and guiding the users through segments relevant to their navigational goals. Our experimental results showed that the segmentation algorithm described in this paper provides up to $80 - 85\%$ success rate in labeling messages. The algorithm is being deployed in a software system to reduce the navigational load of blind students in accessing web-based electronic course materials; however, we note that the techniques are equally applicable for developing web indexing and summarization tools for users with sight.

Categories and Subject Descriptors

H.5.4 [**Information Interfaces and Presentation**]: Hypertext/Hypermedia—*Navigation, user issues*; H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems—*Hypertext navigation and maps*; H.3.3 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Abstracting methods, indexing methods*; K.4.2 [**Computers and Society**]: Social Issues—*Assistive technologies for persons with disabilities*

General Terms

Algorithms, experimentation, human factors, performance

Keywords

Discussion boards, segmentation, navigational aid, assistive technology for blind users

```
buzz proj.           Vander, Ryan       Tue May 25, 2004 9:21 am
 Re: buzz proj.      True, Thomas       Thu May 27, 2004 7:53 pm
  Re: buzz proj.     Vander, Ryan       Sat May 29, 2004 2:08 pm
   Re: buzz proj.    Grain, Robert      Sun May 30, 2004 6:10 pm
    Re: buzz proj.   Vander, Ryan       Sun May 30, 2004 10:23 pm
Assignment 4         Rodriguez, Luisa   Thu May 27, 2004 3:04 pm
Report for Assig. 4  True, Thomas       Thu May 27, 2004 7:57 pm
 Re: Report for Assig. 4  Candan, Kasim Mon May 31, 2004 12:07 am
Assignment #4        Atilla, John       Fri May 28, 2004 10:41 pm
 Re: Assignment #4   Candan, Kasim      Mon May 31, 2004 12:19 am
Questions on #4      Roosewelt, Daniel  Sat May 29, 2004 11:00 pm
 Re: Questions on #4 Candan, Kasim      Mon May 31, 2004 12:23 am
   Re: Questions on #4   Ray, Luisa     Mon May 31, 2004 10:34 pm
     Re: Questions on #4   Home, Chris  Tue Jun 1, 2004 12:23 am
Report Length       True, Thomas        Tue Jun 1, 2004 11:39 am
 Re: Report Length  Candan, Kasim       Wed Jun 2, 2004 1:39 am
Assignment # 4      Bird, Sarah         Tue Jun 1, 2004 9:14 pm
```

Figure 1: A hierarchy of messages posted to a course discussion board: although the *subject headers* of the messages can give some idea about what the postings are about, they provide little information to help differentiate the actual contents of different messages

1. INTRODUCTION

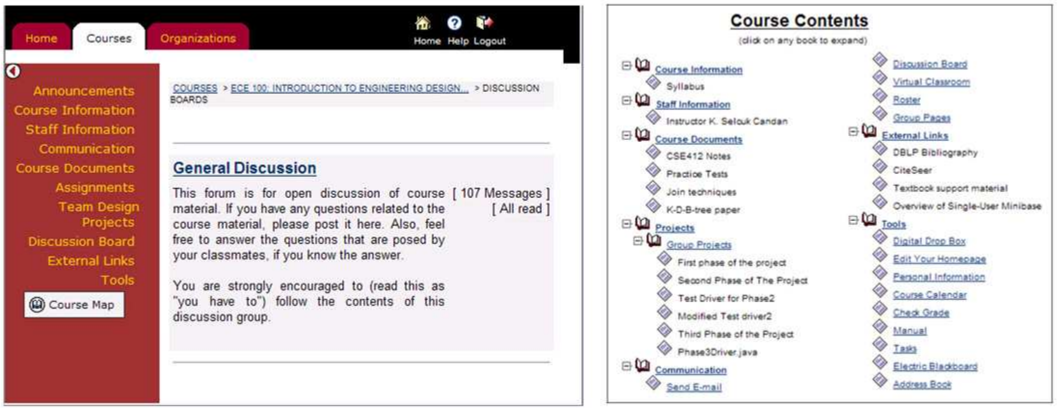
Complex web sites continue to proliferate, as web-based information infrastructures become integral parts of educational, corporate, and e-commerce organizations. Yet, due to the continuously increasing sizes and complexities of these infrastructures, it is also becoming more and more difficult for users to understand and navigate through such sites. The navigation problem is especially critical for users without sight. With the passage of the 508 web accessibility mandate, many companies and federal government agencies are required to follow accessibility guidelines when designing web sites. Such guidelines are very effective when designing mostly static and non-individualized information outlets. However, when

- the material being delivered is information rich yet arbitrarily structured,
- the content is dynamically generated through multiple users' inputs, interactions, and annotations, or
- the users have to follow non-linear, individualized pathways through the material,

the navigational challenge is compounded, even for users with sight. Unfortunately, these characteristics are very common in online course servers and discussion boards.

[*] This work is supported by NSF ITR Grant, ITR-0326544.

WWW 2005, May 10-14, 2005, Chiba, Japan.
ACM 1-59593-046-9/05/0005.



COURSES > ECE 100: INTRODUCTION TO ENGINEERING DESIGN... > DISCUSSION BOARDS

General Discussion

This forum is for open discussion of course material. If you have any questions related to the course material, please post it here. Also, feel free to answer the questions that are posed by your classmates, if you know the answer.

[107 Messages]
[All read]

You are strongly encouraged to (read this as "you have to") follow the contents of this discussion group.

Course Contents

(click on any book to expand)

- Course Information
  - Syllabus
- Staff Information
  - Instructor K. Selcuk Candan
- Course Documents
  - CSE412 Notes
  - Practice Tests
  - Join techniques
  - K-D-B-tree paper
- Projects
  - Group Projects
    - First phase of the project
    - Second Phase of The Project
    - Test Driver for Phase2
    - Modified Test driver2
    - Third Phase of the Project
    - Phase3Driver.java
- Communication
  - Send E-mail
  - Discussion Board
  - Virtual Classroom
  - Roster
  - Group Pages
- External Links
  - DBLP Bibliography
  - CiteSeer
  - Textbook support material
  - Overview of Single-User Minibase
- Tools
  - Digital Drop Box
  - Edit Your Homepage
  - Personal Information
  - Course Calendar
  - Check Grade
  - Manual
  - Tasks
  - Electric Blackboard
  - Address Book

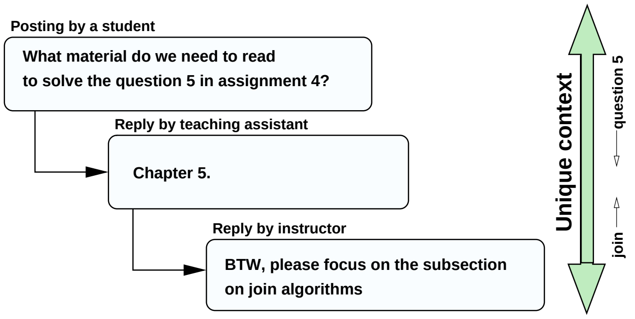
Figure 2: Two sample views from course pages containing announcements, course documents (e.g. lecture notes), course information (e.g. syllabus), assignments, external links, group pages, and *discussion boards*. In this paper, we focus on providing access to discussion boards, such as those included in these samples

1.1 Motivation

Like many others, ASU's educational web site[1] hosts course home pages, containing lecture notes, a syllabus, assignments, project material, course related documents, announcements, external links (links to materials residing in different hosts or different locations in the course server), grades, calendars, group pages, and discussion boards (Figure 2). Some of this content is fixed, meaning that it does not change during a semester (e.g. course syllabi), but majority of the content evolves (e.g. discussion boards) through contributions by the instructors, teaching assistants, and students. Our students without sight emphasized that, although the screen reader software enables them to access the electronic material, they still have to struggle when accessing richly-structured, heterogeneous, and constantly growing content, such as discussion boards (Figure 1). With the aim of reducing the navigational load of blind students, we are developing a software interface, called *iCare-Assistant*, that provides context- and task-dependent navigational guidance when accessing on-line educational materials that are already available for the use of sighted students. State-of-the-art browser-based interfaces [33] and existing navigational helps, such as site maps and visual cues [26], alleviate this load for only sighted users and are generally not applicable to dynamically growing content. Instead, we employ transparent guidance and dynamic adaptation techniques [22, 23] in *iCare-Assistant* to help students without sight. Such dynamic adaptation and guidance requires an understanding of the inherent, but implicit, structures of the content available at the educational web sites. In this paper, we focus on the challenge of identifying coherent information units (or *segments*) in dynamically growing hierarchical content in discussion boards.

1.2 Problem Statement: Topic Segmentation of Dynamically Growing Hierarchical Web Content (Discussion Boards)

Unless a hierarchy corresponds to a well-defined conceptual structure, it does not present information effectively: if the structure is not self-revealing, higher level nodes in the hierarchy cannot direct users to the information available at the lower levels. This is the case for message hierarchies in discussion boards which grow freely through postings of different users at different times: for instance, a posting containing a question may lead to new postings that are not necessarily directly related to the original question. Thus, just by looking at the subject headings or contents of the first few postings in a message hierarchy, one may not be able to guess the actual contents of the replies deeper in the same thread (Figure 1). This complicates the task of navigating within message hierarchies in discussion boards.



What material do we need to read to solve the question 5 in assignment 4?

Chapter 5.

BTW, please focus on the subsection on join algorithms

Figure 3: A chain of three messages: The messages are too short and incomplete for indexing: they obtain their context from their relevant ancestors. Once it is identified that these three messages are within the same context, keywords can be inherited between these messages for proper indexing.

While storing personal (already-read email messages) for reuse, as in Microsoft's *Stuff I've Seen* [24], contextual cues, such as time and author, can be used to search for and present information. However, in a discussion board, where the content is freely growing through multiple users' inputs, interactions, and annotations, such contextual cues may not be enough (Figure 1). In order to provide proper navigational support to users, a guidance system must identify, as precisely as possible, the next possible step(s), based on the current navigational context. When the context changes, the system should adapt to this change by identifying the

[1] `myasucourses.asu.edu`, implemented using the Blackboard software [1]

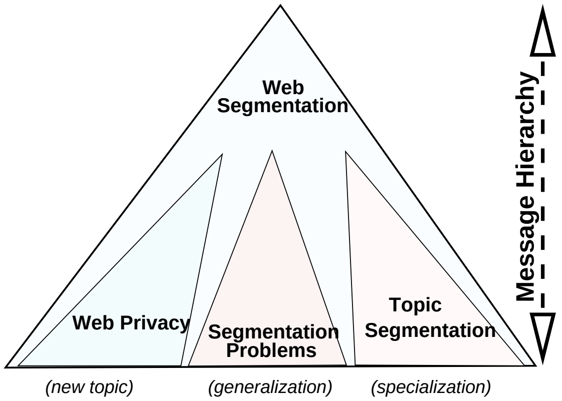


Figure 4: An example showing three types of topic divergences in a message hierarchy: the original discussion theme of "web segmentation" leads to a new discussion topic ("web privacy"), a more general discussion on the topic of the "segmentation problems", and a more specific thread on the 'topic segmentation" issues

most suitable content that has to be brought closer to the user in the navigational space [22, 23]. Such dynamic adaptation of the information space requires an indexing system which can leverage the logical relationships between various contents, such as messages that refer to the same assignment within the same context. Most messages, on the other hand, are too short to be meaningful by themselves, and therefore, they obtain their context from their parents and ancestors (Figure 3). However, as a discussion hierarchy grows through posting of new messages, its content and context will also evolve and possibly diverge from the original posting (Figure 4). Although not all postings will cause a divergence from the initial theme, some of the postings will

- focus on a specific aspect of the original message,
- take the discussion to a more general platform, or
- diverge significantly from the original theme, introducing an entirely new discussion theme.

In a loosely structured environment, where the structure itself is not known, topic distillation [4, 6, 29] and web site summarization [13] algorithms are useful in understanding the underlying structure. In linearly authored (such as text) documents, linear text segmentation techniques [19, 20] can be useful in identifying coherently authored components. However, in freely (and arbitrarily) evolving message hierarchies in discussion boards, the challenge is not to identify how a document is authored, but to discover how the discussion topics have evolved and how they can be *segmented* to identify context (topic) boundaries to facilitate indexing, retrieval, ranking, and presentation of appropriate information units (or segments) to the user.

Thus, the **segmentation problem** within this context can be defined as searching for special nodes − which are the entry points to *new*, *general*, or *specific* topics − within a single hierarchy of dynamically evolving web content (Figure 4). Once the segmentation is completed, each segment can be independently indexed, keyword can be inherited (bottom-up or top-down) based on the generalization and specialization behaviors, and users can be directed to the entry point of the most relevant segment to their current context.

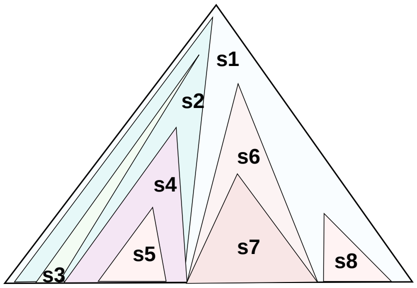


Figure 5: Topic segmentation of a discussion hierarchy

1.3 Contributions of this Paper

In our previous work, we explored web indexing and mining of web information units [31, 32], mining document associations [11, 12, 13], structural mining of hierarchical content [10], and summarization of web sites [13] for sighted people. In this paper, we build on our existing work by developing segmentation (Figure 5) techniques for discovering the topic evolution structures of dynamic and hierarchical web-content, such as discussion boards, for effective indexing and presentation. We develop algorithms for identifying how the topic content of a discussion board evolves through generalizations and specializations as well as introduction of new topics. This knowledge is used in identifying coherent segments of the discussion content. With a precise understanding of the structure of the available discussion content, it could then be possible to fully utilize the context, access history, and user preferences in locating the appropriate discussion segment and presenting it to the user[2]. As described above, these algorithms are being developed to be used in the iCare-Assistant software for blind students in accessing web-based electronic course materials. However, we note that the techniques are equally applicable for developing web summarization tools for users with sight.

2. RELATED WORK

In this section, we present the related work in the domains of topic segmentation, distillation, topic tracking, adaptive hypermedia, video segmentation, adaptive and assistive web technologies, and web community mining.

Topic Segmentation: The idea of topic segmentation has been applied to full-text documents in order to obtain small and coherent documents which can be used as visualization aids [15, 35]. Since the focus of this research has been the segmentation of text documents, underlying techniques have been borrowed from the text segmentation [19, 20] literature. The main difference between the text segmentation and discussion board segmentation is that, while text documents usually present a coherent (*authored*) linear structure that can be exploited for segmentation, discussion boards *evolve* through (mostly short) postings by many contributors. Thus, linear text segmentation [19, 20] techniques are not directly applicable in this domain.

[2] The segment indexing and presentation techniques are outside of the scope of this paper.

Topic Distillation: Hypermedia has two aspects: content and structural information. Web structures can be used as clues while indexing and presenting content. Various techniques have been proposed to use the web structure in identifying document associations, such as the companion and co-citation algorithms proposed by Dean and Henzinger [21]. One approach to organizing web query results based on available web structure is *topic distillation* proposed in [29]. This technique organizes topic spaces as a smaller set of hub and authoritative pages, and thus, it provides an effective mean for summarizing query results. [4] improved the basic topic distillation algorithm presented in [17] through additional heuristics. [6] further considers page fanout in propagating scores. Topic distillation has been used by many search engines, including Google, IBM Clever [17], and TOPIC [9]. Note that topic distillation [4, 6, 29] could be a natural choice for summarization purposes. However, these techniques are usually general purpose and ignore the special hierarchical and dynamic structure of the web content, such as discussion boards. The techniques, we develop in this paper, on the other hand, exploit these two inherent features to establish the underlying segmentation framework.

Topic Tracking: Like the topic distillation work described above, topic detection and tracking (TDT) research [3, 5, 37, 40], which mainly focuses on detecting and tracking events in streaming news data, is related to the work presented in this paper. TDT systems monitor continuously updated news stories and try to detect the first occurrence of a new story; i.e., an event significantly different from those news events seen before. To detect the first story, current TDT systems compare a new document with the past documents and make a decision regarding the novelty of the story based on the content-based similarity values. For example, the method proposed in [5] is based on an incremental TF-IDF model, and it involves segmentation of documents to locate all stories on a previously unseen (new) event in a stream of news stories. In contrast, the naturally evolving nature of discussion threads and the need for fine-granularity segment boundary identification make the problem of topic segmentation significantly harder than the new-event detection problem addressed by the TDT technologies.

Adaptive Hypermedia: Adaptive hypermedia is a rich research field that dates back to the early 1990s [9]. Adaptive hypermedia uses two different but complementary methods, namely adaptive presentation and adaptive navigation [9]. Adaptive presentation is manipulation of content fragments in a hypertext document. Order of fragments can be changed, or fragments can be made invisible or less visible within a page. Stretchtexts, where text fragments can be stretched or shrunk on the basis of user interests, are also used. Adaptive navigation, on the other hand, is the manipulation of links. Direct guidance, link sorting, link hiding, link annotation, link generation, and map adaptation are the techniques used. Detailed discussion of all these approaches, both for adaptive presentation and adaptive navigation, can be found in [7, 8, 9, 14, 16]. Researchers in the AI community have developed web navigation tour guides, such as WebWatcher [28]. WebWatcher utilizes user access patterns in a particular web site to recommend users proper navigation paths for a given topic. User access patterns can also be incorporated into the algorithms we present in this paper. [30] presents a technique for constructing multi-granularity and topic-focused site maps. Their technique can help in visualizing the topology of the web site; thus, it supports navigation. Nonetheless, most of these approaches exploit visual cues as they are designed to help sighted individuals.

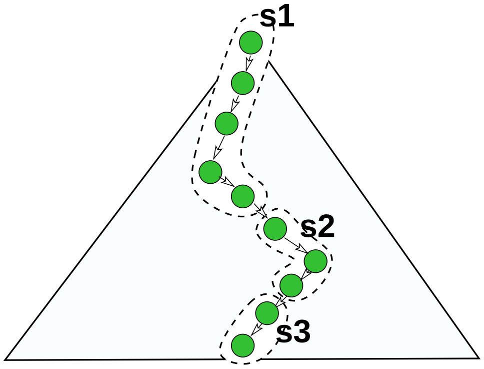


Figure 6: Special case: segmenting a single root-to-leaf chain

Video Segmentation: Video segmentation literature [27, 34, 36, 37] is also relevant to the work presented in this paper. In video, shot (or segment) boundaries are usually detected by comparing various features of consecutive frames or neighborhoods of frames to identify major content changes [27]. Thus, spatio-temporal continuity of common features (e.g. objects, color histograms) shared between two consecutive frames increase the likelihood that these two frames are part of a single coherent segment. As we discuss in the next section, in messaging systems, the varying (but short) sizes of the messages and arbitrarily used (intended, forgotten, or implied) quotations from the ancestor messages further complicate the detection of segment boundaries.

Adaptive and Assistive Technologies: Technologies relied upon by the users with visual impairments include screen readers, screen magnifiers, voice recognition software, hypermedia to hypertext transformers, and refreshable Braille displays. State-of-the-art browser-based interfaces [33] and navigational helps [26] mostly rely on visual guidance, which is not useful for users who are blind. In this paper, we do not focus on specific adaptive technologies exploited to make educational sites accessible [22, 23]. Instead, we present the underlying enabling technology of topic segmentation for discussion boards.

Web Community Mining: Web communities, such as discussion boards and Usenet, are places where people freely participate in discussions. Even though web communities contain a lot of human knowledge, many search engines which have been successful for general purpose web data do not apply well because they ignore inherent structures of the web communities and furthermore postings are usually short. [38] creates a specialized ranking function for Usenet by using linear regression and support vector machine techniques. Their approach is based on metadata, such as prior knowledge about the message author or the depth of the message. They do not address short message problem. [39] suggests a method to extract information from web discussion boards and email archives by summarizing threads. To extract a thread summary, they use quote and comment relationships, which indicate there are topic bindings between messages.

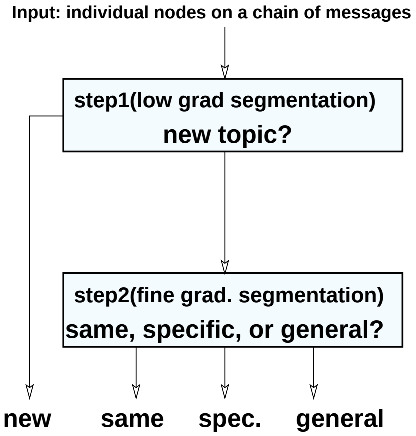

Input: individual nodes on a chain of messages
step1(low grad segmentation)
new topic?
step2(fine grad. segmentation)
same, specific, or general?
new
same
spec.
general

Figure 7: Two step segmentation of a message

3. SEGMENTATION OF MESSAGE HIERARCHIES

Once a hierarchy of messages is segmented as in Figure 5, each segment can be treated as an atomic entity (for instance, if keyword vectors are used for indexing, such vectors can be extracted for the collection of messages in a given segment), or a *key* message (for instance, the first message in the segment) can be chosen to represent the segment. Similar techniques are used in shot (or segment) identification and indexing [18, 25, 27, 34, 36] in linear video streams. Therefore, before tackling the problem of segmentation of a hierarchy of messages, we first focus on the special case of segmentation of a single root-to-leaf chain (Figure 6) in the hierarchy. In Section 3.2, we will extend this for the general case to the segmentation of entire hierarchies.

3.1 Segmenting a Single Message Chain

The approach of segmentation of a sequence of documents was effectively utilized in detecting news stories about a previously unseen event in a stream of news stories [5]. The segmentation technique used (comparing each new document with all, or a carefully selected few, of the previously seen documents to identify in which cluster they belong) is good when the goal is to identify if a document is content-wise similar to a previously seen group of documents. However, when the required segmentation is of finer quality, as in trying to identify whether the topic of the current message is more specific or more general than the topic of its ancestors, such a comparison is not sufficient. Therefore, in this paper, we propose a two-step approach to segmentation (Figure 7): we process the message nodes in a chain in a top-down manner; for each node,

- first, we perform a *low-granularity* segmentation to identify whether the message is of an unrelated topic (relative to the postings immediately before it in the same thread) or not;
- in the second step, if the message is identified to be similar to the previous messages, a *higher-granularity* segmentation process, which tries to determine whether the message is more *specific* or more *general* than the previous messages, is carried out.

In this section, we discuss these two steps in detail.

Message 1: Quotation (left at the bottom of the message) does not provide context

```
Thanks a lot...BTW, is it possible that
two different nodes have the same ad-labels? I found in the
data file produced by JK's code, there exists two different
nodes with the same ad-labels and different ses-labels!

Quote> At this stage there is no planner; we simply pick a
Quote> sequence of MI joins, such that at each stage results
Quote> from one operator join with results from another one.
```

Message 2: Quotation (this time intentionally kept at the top, before the reply text) is included to provide context to the reply

```
Quote> two different nodes have the same ad-labels? I found in the
Quote> data file produced by JK's code, there exists two different
Quote> nodes with the same ad-labels and different ses-labels!

..this is curious... can you please give us more details
regarding this case?
```

Figure 8: The quotations in Message 1 do not provide a common context, whereas the quotations in Message 2 do provide a common context

3.1.1 Step I: Identifying New Topic Boundaries

In this step, we identify whether the current message is sufficiently different from the previous postings in the same thread to be marked as a *new* topic. Unlike in a stream of news documents [3, 5, 37], where different news may be interleaved in a given sequence, in a given thread of a discussion board, there is a natural tendency of maintaining the same topic because most postings are replies to previous ones. Thus, unlike the previous work on TDT, a new node does not need to (and cannot) be compared to all its ancestors, but has to be compared to its immediate parent (or an immediate sequence of ancestors) as it is (they are) causally closest to the current node.

A similar approach of comparing the features of a frame locally with its immediate predecessors works well in identifying shot boundaries in video streams [27]. When comparing two consecutive video frames, any of the common features (e.g. objects, color histograms) shared between them increases the likelihood that these two frames are part of a single shot (a coherent segment). However, when segmenting discussion threads, there are certain complications:

- First, unlike consecutive video frames that are mostly identical, consecutive messages of the same topic may be of different length, style, and content.
- Secondly, in many messaging systems, original postings are automatically included in replies as quotations; hence, unless quotations are used in a way to strengthen the link between the original message and the reply, they may not highlight a common context (Figure 8).

Thus, keywords in quotations should be treated differently based on the *relevance* of the quotations as determined by their placement in the message; in general, quotations selectively used within the body of a message (Message 2 in Figure 8) are more relevant than the quotations left (potentially forgotten) as a bulk at the end of a message (Message 1 in Figure 8). In this paper, we do not focus on the problem of identifying *selectively-used* quotations, instead we focus on the impact of quotations on the segmentation task.

In general, keywords in quotations can be considered as keywords inherited from the ancestors. By including them in the keyword vector of a message, we can implicitly increase the similarity between the current message and the quoted message. However, keywords in quotations have to be treated differently than the other keywords to prevent undeserved bias. Let us represent each message in the hierarchy as a keyword weight vector $\langle w_1, w_2, \ldots, w_n \rangle$. The weight, w_i, of the keyword k_i is computed using the aggregate frequency of the keyword,

$$freq_i = freq_{i,0} + \sum_{1 \le d \le quot\_depth} imp(d) \times freq_{i,d}$$

where

- $freq_{i,0}$ is the frequency of the keyword k_i in the message excluding the quotations,
- $freq_{i,d}$ is its frequency in d-level quotations (quotations from the parent message, as in Figure 8, are of 1-level), and
- $imp(d)$ is the impact factor of the quotations that are of depth d.

Thus, the contribution of the quotation keywords to the overall frequency of the term varies based on the value of the corresponding impact factor. Impact factors greater than 1 imply that the resulting keyword vector will have a higher similarity to the ancestor from which the quotations have been taken, whereas impact factors less than 1 imply that, although quotations are important, the actual content of the message should be used for determining whether the message is similar to the ancestors or not. Note that, even when the impact factor is only 1, the existence of the quotation keywords in the message gives bias towards increased similarity between the ancestors and the message.

Once a keyword weight vector is computed for a message, the cosine similarity between this vector and the keyword vector of the parent message (or the keyword vector representing the segment being computed so far) can be used to classify the input message as having a *new* topic or being of the *same topic* as that of the parent. Other similarity and distance measures, such as Hellinger distance and Kullback-Leiber divergence are also shown to work well in the TDT domain [3, 5, 37].

3.1.2 Step II: Segmentation based on Specialization/ Generalization

If a message on a given chain is identified to introduce a new topic to the discussion, this message can be used as a segment boundary. On the other hand, if the difference between the message being considered and the earlier messages is not large enough to trigger segmentation, then an initial segmentation is not possible. However, even though a message may not diverge significantly from the initial theme, it may

- focus on a specific aspect of the common theme or
- take the discussion to a more general platform.

Finding such *specialization* and *generalization* boundaries is also important because understanding when a discussion topic diverges helps both with indexing (by choosing the right keyword weights for the given segment) as well as guiding the user (without sight) to the most appropriate entry point within a discussion. Therefore, in this step, among the parent/child messages that are identified to be of the same topic, we need to detect specialization and generalization boundaries. For this purpose, we first need to define the terms *specialization* and *generalization*.

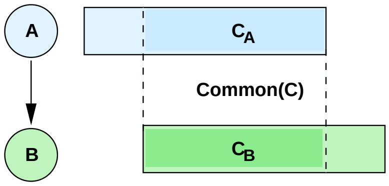


Figure 9: Visual representation of the *contents* of two (parent/child) messages of the same topic: the two messages share a common base (a common context), but they also have their own content

In general, as shown in Figure 9, given two messages, A and B, of the same topic, they will have a common base, while both messages will also have their own content, different from their common base. If the common base, C, of these two messages can be identified, then the degree of *specialization* can be defined as

$$spec(A, B) = 1 - similarity(A, C_B),$$

where C_B is the content in message B corresponding to the common base with A. Intuitively, given two messages, A and B, that are already identified as being of the same topic, if the original message, A, is not similar to the common base of the two messages, it means that the common base is a small part of the original message; i.e., the new message *specializes* within the original message.

Similarly, given the common base between A and B, the degree of *generalization* can be defined as

$$gen(A, B) = 1 - similarity(B, C_A),$$

where C_A is the content in message A corresponding to the common base with B. Again, intuitively, given A and B of the same topic, if the new message, B, is not similar to the common base of the two messages, this would mean that the common base is a small part of the new message; i.e., the new message *generalizes* on the original message.

Unfortunately, in practice, identifying the common base of two messages and computing the specialization and generalization degrees by comparing the two messages to this common base are not trivial tasks. We use the quotations from previous messages to help us with this process. Thus, we fragment each message on a discussion board into zero or more *anchored* parts and a *free* part. An anchored part of a message is composed of the quotation messages from the parent and ancestors as well as the parts of the message identified to be replies to these quotations. For instance, Message 2 in Figure 8 is composed of a quotation-reply pair; in a sense, the quotation message is a context-providing pointer to the ancestor, which can be used to improve the accuracy of segmentation. The free part of a message is the part which is not immediately associated with the parent or ancestor quotations.

For the anchored parts of a message, if quotations from the parent or ancestors are used as context-providing pointers, then the degree of *specialization* or *generalization* should be defined within the associated context. Taking this into account, we define the degree of *generalization* as

$$\frac{1}{N}\left(n_{free} \times gen(D_{par}, d_{free}) + \sum_{d_i \in anchored} n_{anch,i} \times gen(q_i, d_i)\right)$$

where

- N is the number of keywords in the message,
- d_{free} is the free part of the message and n_{free} is the number of keywords in this part,
- q_i, d_i is the i^{th} anchored quotation-reply pair and $n_{anch,i}$ is the number of keywords in this pair, and
- D_{par} is the parent message.

Note that, while the free part of the message is compared directly against the parent (assuming that the parent, which is of the same topic, provides the context), the anchored components are compared against the corresponding context as highlighted by the quotation. The degree of *specialization* is defined similarly:

$$\frac{1}{N}\left(n_{free} \times spec(D_{par}, d_{free}) + \sum_{d_i \in anchored} n_{anch,i} \times spec(q_i, d_i)\right),$$

Finally, once the degrees of generalization and specialization are computed for given two messages, A and B, if $gen(A, B) > \Theta_g$, for a given generalization threshold, Θ_g, then B is marked as a generalization boundary. When this is not the case, if $spec(A, B) > \Theta_s$, for a given specialization threshold, Θ_s, then B is marked as used as a specialization boundary within the same topic. If neither of these cases is true, then B and A are said to be in the same topic segment. Note that, although we do not elaborate in this paper, these threshold values need to be set through a *learning* process which identifies proper thresholds based on a given training sample. Nevertheless, in Section 4, we experimentally evaluate the effects of different threshold values on the performance of the algorithm.

3.2 Segmenting a Hierarchy of Messages

Once we establish the techniques for segmenting a root-to-leaf chain in a given message hierarchy, extending these for achieving the segmentation of the entire hierarchy is straight-forward. Since two separate replies to a single message are independently created from each other, they cannot be marked to be of the same topic, unless they are independently identified to be of the same topic as that of their common parent. Thus, the two-step segmentation process described above can be repeated in a top-down fashion, following each chain of the hierarchy independently. Finally, each connected component of the tree, not split with segment boundaries, is marked as an atomic segment and indexed separately, while the *specialization* and *generalization* information is used to identify how keywords are inherited between ancestors and descendents[3]. The common ancestor of all nodes in a given segment is identified as the *entry point* of the segment and used in guiding users.

Table 1: Weights used to measure undifferentiated, low-only, and differentiated errors

	Undiff.				Low-only				Diff.			
	N	S	SF	SG	N	S	SF	SG	N	S	SF	SG
N	0	1	1	1	0	1	1	1	0	1	1	1
S	1	0	1	1	1	0	0	0	1	0	0.5	0.5
SF	1	1	0	1	1	0	0	0	1	0.5	0	0.5
SG	1	1	1	0	1	0	0	0	1	0.5	0.5	0

Table 2: The weighted success rate for the proposed algorithm is greater than 79%, even for the undifferentiated scheme, where all errors are counted

Success rate (%)		
Undiff.	Low-only	Diff.
79.06%	87.31%	83.19%

4. EXPERIMENTAL EVALUATION

In order to evaluate the effectiveness of the segmentation techniques presented in this paper, we performed a user study and compared the segmentation feedback provided by assessors of a discussion board with the segmentation results obtained by the proposed algorithm.

Setup: For the evaluations presented here, due to the diversity of its postings and message hierarchies, we used the movie message board available at [2] as the message data source. We randomly selected

- 20 discussion threads, with
- a total of 368 messages,
- average thread depth of 12.45,
- average quotation depth of 1.3 (86% of the total of 5241 quotations are from the parent)

from this source and asked 5 users to assess each message to label it with N for new topic, S for same topic as the parent, SF for specialization (or focussing), or SG for generalization. Given all manual labelings from multiple assessors, we took the majority label to denote the page's relationship with its parent. We then compared these manual labeling results with the labels assigned by the proposed automated segmentation algorithm (which took only **560ms** to segment the given 20 threads). In this section, we report the results when the threshold for detecting *new* segment boundaries is set to 0.35, generalization threshold, Θ_g, is set to 0.6, and the specialization threshold, Θ_s, is set to 0.8 (we discuss the effects of varying these thresholds later in the section). Also, for the results presented here, the impact factor for the parent quotations ($d = 1$) is $imp(d) = \frac{1}{2}$ (we discuss the effect of different impact factor values in the later section).

Evaluation criteria: In order to observe the effectiveness of the proposed algorithms, we computed a labeling *success rate* (or precision),

$$success\_rate = \frac{\sum_{m \in messages} 1 - error\_weight(m)}{number\ of\ messages} \times 100,$$

where error weights are used to account for *gravity* of the error in the computed success rate. We experimented with three different schemes as shown in Table 1:

- *Undifferentiated weights:* Weights in first partition of the table mark all errors with the same (maximum) error weight, independent of the type of error.
- *Low-only weights:* Weights in second partition in the table only count errors in the first, low-granularity, step of the algorithm; i.e., only

[3] The details of the indexing process of the segments are out of the scope of this paper.

Table 3: Distribution of various types of errors

Alg.\ User	**New**-u	**Same**-u	**Spec.**-u	**Gen.**-u	*Tot.*
New-a	–	31.0	1.4	7.0	*39.4*
Same-a	14.1	–	16.9	11.3	*42.3*
Spec.-a	0.0	4.2	–	0.0	*4.2*
Gen.-a	7.0	5.6	1.4	–	*14.1*
Total	*21.1*	*40.9*	*19.7*	*18.3*	*100*

Table 4: User labelings for the 368 messages in the randomly selected 20 threads

New	Same	Spec.	Gen.	No Majority (unlabeled)	Tot.
58	206	39	36	29	368

  - those pages that are marked erroneously as being of a *new topic* or
  - those that should have been marked as a *new topic*, but not marked as such

  count towards the error rate.

- *Differentiated weights:* Weights in third partition in the table penalize different error types differently. More specifically, errors within the high-granularity group (S, SF, and SG) are marked half as costly as errors across the low-granularity segmentation.

Table 2 shows the weighted success rates observed in the experiments.

Undifferentiated success rate: Based on the user study, we observed that the undifferentiated success rate, where all errors are penalized with the maximum weight without distinguishing between the types of errors, was around 79.06% (first column in Table 2).

Low-only success rate: On the other hand, when we focus on only the errors in the first, low-granularity, step of the algorithm, we observed that the success rate jumped to 87.31% (second column in Table 2).

Differentiated success rate: When a differential penalty scheme (where errors within the high-granularity group $-$ S, SF, and SG $-$ are marked half as costly as errors across the low-granularity segmentation between same and new topics) is used, the success rate was 83.19% (last column in Table 2).

Distribution of the errors: Table 3 provides a detailed tally of the types of errors (around 20% of all labelings as described above) observed during the user study. In this table, the columns correspond to the labelings chosen by the users, while the rows correspond to those assigned by the proposed algorithm.

As can be seen by studying the last row of the table, which shows the aggregate number of the errors made by the proposed algorithm for each user labeling, the greatest percentage (40.9%) of labeling errors is due to messages that are marked *same* by the users. In fact, the biggest single contributor to the number of errors is the set of *same topic* messages that are labeled as *new* by the algorithm (31% of all errors). In the last column of the table, which shows how the errors are distributed among labeling of the algorithm, we see that 42% of all errors are due to messages that are incorrectly marked *same*, whereas around 40% of the errors are due to those that are incorrectly marked as *new.* The total contribution of specialization and generalization errors to the overall rate of the error is less than 20%.

Table 5: Success rates for individual labelings

Success rate for labelings				
New	Same	Spec.	Gen.	*Overall*
0.74	0.86	0.64	0.64	*0.79*

Table 6: The impact of quotations on the labeling performance

	Success rate (%)		
Quot. weights	Undiff.	Low-only	Diff.
Off	72.57%	85.25%	78.90%
On	79.06%	87.31%	83.19%

Note that, since the distribution of labels provided by the users is not uniform (Table 4), the impact of different types of errors on the overall success rate varies. Table 5 presents success rates achieved by the proposed algorithm for each label. The success rate achieved for those messages labeled *new* by the users is around 74%. The success rate is as high as 86% for detecting messages that stay within the *same* topic. The fine granularity segmentation success rate in the second phase is around 64%. As can be seen from the **spec.**-u and **gen**-u columns in Table 3, most of the errors in the second phase of the algorithm are due to messages that are marked *same topic* by the algorithm but further classified into *specialization* and *generalization* categories by the users. This shows that, while the human assessors can differentiate fine topic distinctions better, the proposed algorithm may *conservatively* classify messages to be of the *same* topic to prevent over-segmentation. The overall (undifferentiated) success rate is close to 80%, as described earlier.

Effect of quotations: In order to observe the impact of the quotations on the performance of the segmentation algorithm, we calculated how the success rates changed when the context-sensitive weighting techniques proposed in this paper were turned off. When the quotations were not treated specially, the number of errors in the first step of the algorithm increased 11%, from 43 to 48 erroneous labelings. On the other hand, the total number of errors (including both phases of the algorithm) increased 30%, from 71 to 93, showing that especially the fine-granularity differentiation required in the second phase benefits significantly from the way the proposed algorithm uses quotations for context-sensitive weighting (Table 6).

In Table 3, we saw that 31% of the all errors were due to the set of *same topic* messages that were labeled as *new* by the algorithm. In order to see whether using a different impact factor formulation would improve this situation, we tried impact factors with different characteristics. A selection of the low-granularity (*same* versus *new*) labeling errors are reported in Table 7. The first row of this table corresponds to the results presented so far. The following rows shows the results obtained when the impact factors were set such that the resulting keyword vector would have a higher similarity to the ancestor from which the quotations have been taken. The results show that, indeed, the number of *same topic* errors drops when the impact of the keywords in the quotations increases. However, this is accompanied with a significant jump in the number of *new* messages that are labeled as *same*, reducing the overall success rate as shown in the last column of Table 7. In fact, between the two extremes (first and last rows) in the table, *new* message identification ($30 - 15 = 15$) is more sensitive to the weight of

Table 7: The effect of quotation impact factors on the low -granularity labeling performance

$imp(d)$ for $d = 1$	same$\xrightarrow{err}$ new	new $\xrightarrow{err}$ same	*undiff. succ.*
0.5	28	15	79.0%
1	23	18	77.6%
1.5	24	26	77.6%
2	22	30	76.4%

Table 8: Effects of different Θ_g and Θ_s thresholds

	0.0	**0.25**	**0.5**	**0.75**	**1.0**	*Exp.*
Undiff.	0.21	0.33	0.56	0.77	0.74	*0.79*
Low-only.	0.87	0.87	0.87	0.87	0.87	*0.87*
Diff.	0.51	0.57	0.69	0.82	0.77	*0.83*

quotations than *same* message identification ($28 - 22 = 6$). Thus, overweighting quotations does not help the overall success rate.

The effect of threshold values: Finally, Table 8 shows the effect of various Θ_g and Θ_s values on the final success rate. As expected (since it is insensitive to the fine-granularity segmentation), the low-only success is independent of the values of Θ_g and Θ_s thresholds. Note that neither too small nor too large values are good for proper segmentation. As we mentioned earlier, threshold values need to be set through a *machine learning* process which identifies proper values based on a given training sample.

5. CONCLUSIONS

Message threads evolve with new postings as new messages may focus on or diverge from the original theme of the thread. In this paper, we presented algorithms for identifying how the hierarchical content of a discussion board grows through generalizations and specializations. This knowledge can be used in segmenting the message hierarchy into coherent units to facilitate indexing, retrieval, and ranking, as well as in guiding users through *segments* that are relevant for their navigational goals. The segmentation algorithms are being deployed in a software system, called iCare-Assistant, which aims at reducing the navigational load for blind students in accessing web-based electronic course materials through an unobtrusive, task-oriented, and individualized delivery interface. However, we note that the techniques are equally applicable for developing web summarization tools for users with sight.

6. ACKNOWLEDGMENTS

We would like to thank our students without sight, Kenneth Spector and David Paul, for their feedbacks and suggestions on the usability of iCare-Assistant; Terri Hedgpeth for her expertise in assistive technology; and Shibo Wu, Lina Peng, Yan Qi, and Ping Lin for their help in various phases of the implementation and evaluation of our work.

7. REFERENCES

[1] Blackboard. http://www.blackboard.com.

[2] Movie message board. http://www.hundland.com/movieboard.mv.

[3] J. Allan, C. Wade, and A. Bolivar. Retrieval and novelty detection at the sentence level. In *26th annual ACM SIGIR Conference*, Toronto, Canada, July 2003.

[4] K. Bharat and M. Henzinger. Improved algorithms for topic distillation in a hyperlinked environment. In *21st Annual International ACM SIGIR Conference*, pages 104–111, Melbourne, Australia, August 1998.

[5] T. Brants, F. Chen, and A. Farahat. A system for new event detection. In *26th annual ACM SIGIR Conference*, Toronto, Canada, July 2003.

[6] S. Brin and L. Page. The anatomy of a large scale hypertextual web search engine. In *7th Int. WWW Conf.*, 1995.

[7] P. Brusilovsky. Methods and techniques of adaptive hypermedia. *User Modeling and User-Adapted Interaction*, 6(2-3):87–129, 1996.

[8] P. Brusilovsky. Efficient techniques for adaptive hypermedia. *Intelligent Hypertext: Advanced techniques for the World Wide Web, Lecture Notes in Computer Science, Springer Verlag*, 1326:12–30, 1997.

[9] P. Brusilovsky. Adaptive hypermedia. *User Modeling and User-Adapted Interaction*, 11(1), 2001.

[10] K. S. Candan, J. W. Kim, H. Liu, and R. Suvarna. Structure-based mining of hierarchical media data, meta-data, and ontologies. In *5th Workshop on Multimedia Data Mining in conjunction with the ACM Conference on Knowledge Discovery and Data Mining*, Seattle, WA, August 2004.

[11] K. S. Candan and W.-S. Li. Using random walks for mining web document associations. In *Pacific Asian Conf. on Knowledge Discovery and Data Mining*, 2000.

[12] K. S. Candan and W.-S. Li. Discovering web document associations for web site summarization. In *DaWaK*, pages 152–161, 2001.

[13] K. S. Candan and W.-S. Li. Reasoning for web document associations and its applications in site map construction. *Int. Journal of Data and Knowledge Engineering*, 2002.

[14] M. Cannataro, A. Cuzzocrea, and A. Pugliese. A probabilistic approach to model adaptive hypermedia systems. In *1st International Conference on Web Dynamics (in conjunction with the 8th Int. Conference on Database Theory)*, pages 12–30, London, UK, January 2001.

[15] C. Caracciolo, W. van Hage, and M. de Rijke. Towards topic driven access to full text documents. In *European Digital Library Conferences*, 2004.

[16] T. Cavanaugh. The need for assistive technology in educational technology. *Educational Technology Review*, 10(1), 2002.

[17] S. Chakrabarti, B. Dom, P. Raghavan, S. Rajagopalan, D. Gibson, and J. Kleinberg. Automatic resource compilation by analyzing hyperlink structure and associated text. In *7th World Wide Web Conference*, pages 65–74, Brisbane, Queensland, Australia, April 1998.

[18] S.-F. Chang, W. Chen, H. J. Meng, H. Sundaram, and D. Zhong. VideoQ: An automated content based video search system using visual cues. In *ACM Multimedia Conf.*, Seattle, WA, 1997.

[19] F. Choi. Advances in independent linear text segmentation. In *ANLP-NAACL-00*, 2000.

[20] F. Choi. Linear text segmentation. In *CLUK3*, 2000.

[21] J. Dean and M. Henzinger. Finding related pages in

the World Wide Web. In *8th World Wide Web Conference*, Toronto, Canada, May 1999.

[22] M. E. Donderler, K. S. Candan, S. Wu, L. Peng, and J. W. Kim. Adaptive electronic course content delivery for students who are blind. In *Demonstration at ACM Conf. on Computers and Accessibility (ASSETS04)*, 2004.

[23] M. E. Donderler, L. Peng, and K. S. Candan. Adaptive content delivery to assist blind students in accessing course materials. In *6th Annual Accessing Higher Ground Conference: Assistive Technology and Accessible Media in Higher Education*, Boulder, CO, November 2003.

[24] S. T. Dumais, E. Cutrell, J. J. Cadiz, G. Jancke, R. Sarin, and D. C. Robbins. Stuff i've seen: A system for personal information retrieval and re-use. In *Proceedings of SIGIR*, 2003.

[25] A. Hampapur, R. Jain, and T. E. Weymouth. Indexing in video databases. In *SPIE/IS&T Proceedings on Storage and Retrieval in Image and Video Databases*, volume 2420, pages 292–306, San Jose, CA, 1995.

[26] I. Herman, G. Melancon, and M. S. Marshall. Graph visualization and navigation in information visualisation: a survey. *IEEE TVCG*, 6(1):24–43, 2000.

[27] L. A. R. J. S. Boreczky. Comparison of video shot boundary detection techniques. In *SPIE/IS&T Intern. Symposium Electronic Imaging*, volume 2664, pages 170–179, 1996.

[28] T. Joachims, D. Freitag, and T. Mitchell. Webwatcher: A tour guide for the world wide web. In *International Joint Conference on Artificial Intelligence*, August 1999.

[29] J. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46:604–632, 1999.

[30] W.-S. Li, N. F. Ayan, O. Kolak, Q. Vu, H. Takano, and H. Shimamura. Constructing multi-granular and topic-focused web site maps. In *10th World Wide Web Conference*, Hong Kong, China, May 2001.

[31] W.-S. Li, K. S. Candan, Q. Vu, and D. Agrawal. Query relaxation by structure and semantics for retrieval of logical web documents. *IEEE TKDE*, 2001.

[32] W.-S. Li, K. S. Candan, Q. Vu, and D. Agrawal. Retrieving and organizing web pages by information unit. In *WWW Conf.*, pages 230–244, 2001.

[33] H. Lieberman. Letizia: An agent that assists web browsing. In *14th Int. Joint Conf. on Artificial Intelligence*, Montreal, Canada, 1995.

[34] N. V. Patel and I. K. Sethi. Video shot detection and characterization for video databases. *Pattern Recognition*, 30:583–592, 1997.

[35] J. Reynar. Topic segmentation: Algorithms and applications. In *PhD thesis, University of Pennsylvania*, 1998.

[36] H. Zhang, A. Kankanhalli, and S. Somaliar. Automatic partitioning of full-motion video. *Multimedia Systems*, 1:10–28, 1993.

[37] Y. Zhang, J. Callan, and T. Minka. Novelty and redundancy detection in adaptive filtering. In *25th annual ACM SIGIR Conference*, Tampere, Finland, August 2002.

[38] W. Xi, J.Lind, and E. Brill. Learning Effective Ranking Functions for Newsgroup Search. In *27th annual ACM SIGIR Conference*, Sheffield, UK, July 2004.

[39] A. Murakami and K. Nagao and K. Takeda. Discussion Mining: Knowledge Discovery from Online Discussion Records. The First NLP and XML Workshop, Tokyo, Japan, November 2001.

[40] J. Kleinberg. Bursty and hierarchical structure in streams. In *8th annual ACM SIGKDD Conference*, Edmonton, Canada, July 2002.