Prototype-Matching System for Allocating Conference Papers

Kloptchenko Antonina *, Back Barbro*, Vanharanta Hannu, Toivonen Jarmo**, Visa Ari**

*Turku Center for Computer Science, Åbo Akademi University, Turku, Finland
Pori School of Technology and Economics, Pori, Finland
**Tampere University of Technology, Tampere, Finland
Phone*: (358)2215-3319, fax*: (358)2215-4809
Antonina.Kloptchenko@abo.fi, Barbro.Back@abo.fi, Hannu.Vanharanta@pori.tut.fi,
Jarmo.Toivonen@tut.fi, Ari.Visa@tut.fi

Abstract

Conferences on applied research require more complicated taxonomy than traditional organization of conferences by tracks. A topic of a paper, submitted to a conference on the applied research and the keywords, outlined by authors can be discussed in more than one proposed conference track. Sorting out the papers submitted to a scientific conference in the proposed categories and tracks is becoming a nontrivial task. Conference organizing committees try to schedule submitted papers very carefully to increase the success rate of the conference. For example, the organizers of The Hawaii International Conference on System Science 2001(HICSS-34) allocated the theme similarities in papers that were submitted into different tracks and identified 6 cross-track themes to schedule them appropriately.

In this paper, we offer a prototype matching system for text retrieval by content and try it out on the HICSS 34 conference proceeding. On the one hand, the system assists the conference organizers to automatically establish semantic similarities among papers and allocate them into common themes. On the other hand, the system assists the attendees to retrieve the papers from the conference proceedings based on their content similarities. A user can take an abstract or a paragraph from an interesting paper, and use it as a prototype query. The information system is based on document preprocessing, "smart" document encoding and prototype-matching clustering of a text collection.

1. Introduction

Traditionally many scientific conferences are organized into tracks. Conferences on applied research have a complicated taxonomy, because of the overlapping borders of applied research fields. The topic of a paper, submitted to a conference on applied research, can belong to several disciplines and be discussed in more than one proposed conference track. On the one hand, the conference organizers have a hard time determining and scheduling overlapping sessions successfully. On the other hand, a conference attendee has a hard time determining which conference sessions are relevant to his/her research interests. He/she needs either to browse the entire conference proceedings to identify interesting papers or to rely on a keyword search, considering keywords as a reflection of the paper content. Authors often use analogous keywords, which can belong to either the same or different tracks to identify the content of the submitted papers. Moreover, the authors and the readers of the scientific articles can represent the same semantics using different words (synonymy) or describe different meanings using words that have various meanings (polysemy) [8]. Sometimes, even experienced readers, such as track chairmen, encounter certain difficulties in the determination of what track the paper should truly belong to.

The amount of text in large conference proceedings requires a new generation of techniques and tools to support scientists in finding critical nuggets of useful knowledge. Quantity, quality and ambiguous structure of available text create many obstacles in working with it. Browsing, searching and organizing text collections turn out to be time consuming and costly procedures. Text mining (TM) methods in form of information retrieval (IR) by content tools strive to assist user information needs. TM is the process of analyzing text to extract information that is useful for particular purposes [24]. While searching text collections for relevant information, users face problems in constructing smart queries because they might not be fully acquainted with the established terminology in a field, or not fully sure about the content of the needed documents. This behavior requires sophisticated IR-by-content tools that could help users to deal with text collections.

Proceedings of the 36th Hawaii International Conference on System Sciences (HICSS'03)
0-7695-1874-5/03 $17.00 © 2002 IEEE

The Hawaii International Conference on System Science (HICSS)[1] is a general-purpose conference that has served the computer society for over three decades. Contrary to many other conferences that have a focus on a specific subject or topic, HICSS addresses a wide range of issues from computer science, computer engineering, and information systems. The objective of HICSS is to "provide a unique environment in which researchers, academicians and practitioners in the information, computer and system sciences can exchange ideas, techniques and applications" [21]. The organizing committee of HICSS tries to build a workshop-like setting at the conference and schedule all the sessions carefully to create a high degree of interaction and discussion among the conference participants. Over the past year it has become clear to organizers that there are similarities or common themes in the papers that were submitted into different tracks. In 2001 the conference organizing committee of HICSS-34 had identified six cross-track themes that united some minitrack from different tracks. The organizers scheduled the papers from those themes very carefully to help conference attendees to participate in all relevant sessions. This case shows that sorting out the papers submitted to a scientific conference in the proposed conference tracks develops into a rather complicated task.

In this paper, we propose a prototype matching system for text retrieval by content. The prototype is a document or a part of it, which is of interest to a particular user. This prototype is matched with an existing document collection. We illustrate the system using a scientific conference collection from The Hawaii International Conference on System Science 2001. On the one hand, the system aims to assist the conference organizers to establish semantic similarities among the papers automatically. On the other hand, the system aims to assist the attendees to retrieve interesting papers from the conference proceeding based on their content similarities. A user can take the whole paper or an abstract from an interesting paper, and use it to construct a smart query. The core of the system is "smart" document encoding on different syntactic levels, and document collection clustering.

The material presented in the remainder of this paper is organized as follows. In Section 2 we review the related work in using text-clustering techniques for organizing text collections to enable information retrieval (IR) by content. In Section 3, we explain the methodology of a prototype-matching system that consists of document encoding, prototype matching and retrieval parts. The prototype-matching part is based on creating histograms for word and sentence levels of every document in a collection. In Section 4 we describe our motivation for creating the prototype matching system for identifying the scientific papers relevant to the user in a conference collection. In Section 5, we give a brief description of HICSS 34 scientific paper collection that we have worked with. Section 6 contains an description of our experiments, and discussion about the results. Finally, in Section 7, we provide some conclusions and suggestions for future work.

[1] http://www.hicss.org/history.pdf

2. Background and related studies

The prototype-matching system that we use in our experiments is an IR by content technique based on document collection clustering. As a good IR system, our system directs a user to the semantically relevant document to satisfy his/her information needs. The characterization of relevance is complex, and thus, for the reasons of efficiency, IR systems use simplistic representation of document content and user information need. Good IR systems, by any mechanism available, should discover this dichotomy [5]. Text collection clustering and term-based approaches for the IR domain have been extensively explored to cope with this complicated task. Below we give a brief overview of the studies made in those approaches.

2.1. Text Clustering for IR by content

Organizing text collections for enabling the retrieval by content [2, 14, 15, 16] and searching [4] can be accomplished by using text collection clustering. Cutting et al. used a clustering technique that supports an iterative searching interface by dynamically scattering a document collection into smaller semantic clusters. A user navigated the document search space by selecting relevant documents among the clusters to regroup the results [4]. Anick and Vaithyanathan exploited document clustering and paraphrasing of term occurrence for document retrieval by content [2]. Merkl and Schweighofer used a different approach for detection of the similarities between documents in organized legal text corpora to enable document retrieval by content [16]. They combined a vector space model, cluster analysis and Self-Organizing Maps (SOM) to organize the legal text corpora as the hypertext and knowledge base of descriptors, probabilistic context-sensitive rules and meta-rules of legal concepts. Lee and Yang presented a SOM-based clustering approach based on word co-occurrences for IR on a Chinese corpus from the web [14]. SOM is a general unsupervised tool popular for clustering and visualization of very large document collections [11]. SOM organizes high-dimensional input data so that similar inputs are mapped close to each other. The

WebSom system is based on SOM clustering and allows browsing and retrieval of the resulting matching list, allowing the user to navigate a multi-level search of text collection [12]. Lin et al. explored the potentials of the SOM semantic map as a retrieval interface for an online bibliographic system [15]. In a majority of the above-mentioned algorithms for IR by content, the user participates actively in the clustering and navigating processes, controlling the fulfillment of his/her information needs.

The effective representation of text, the determination of similarity, and the high dimensionality of document collections are primary challenges in text collection clustering for retrieval by content. Effective solutions to these challenges are discussed in [2, 8, 18], and [19]. Robustness and expendability are important for practical use of IR by content methods. The majority of content-based retrieval systems are based on computational linguistic approaches and linguistic knowledge about the text collection. Hatzivassiloglou et al. noticed that linguistically motivated features in conjunction with full word vectors increase the overall clustering performance [9]. Miike et al. developed a Japanese full-text retrieval system that analyzes text and enables the user to generate an abstract interactively [17]. The system was based on linguistic knowledge and clues, such as idiomatic expressions and was domain independent but required a dictionary of 60,000 entries for morphological analyzes of sentences.

2.2. Term-based approaches for IR

Other commonly used approaches for IR by content are based on user-defined *term-based* methods, such as keywords [3, 6, 10, 20], indexing [13, 23], or mark-ups [1]. Conversely, some valuable information hidden in the documents, which is not outlined by manually or automatically chosen keywords, indexes and markups, cannot be retrieved.

Keyword based clustering approaches have been studied by Sparck-Jones in [20]. Keywords from the Dewey decimal classification in content of books in the United States characterize text well but lack in accuracy [6]. Chien used Patricia tree for extracting the assigned by the author keywords and characterizing the content [3]. Jo assigned categorical substantial weights for informative, functional and alien keywords for text categorization [10].

C. van Rijsbergen studied the classical indexing IR approaches [22]. Lawrence created a full-text index of scientific literature on the web aiming at dissemination, retrieval and accessibility of the scientific literature [13]. The authors used the standard practice of indexing by building hash-table of words (inverted index) that contained a compressed version of the word and a pointer to a block of a record file corresponding to the positions in a matching document.

Markup tells how to display the material, rather than identifying what the material is. User-defined markups help to structure and categorize hypertext documents [1]. Keywords, headings and indexes can be used to mark and to create tags to the interesting document, i.e. flexible markups.

We designed our prototype-matching language independent clustering system to enable IR by content. It differs from the methods mentioned above because it does not focus on word co-occurrences [14], and does not create a high dimensional vector space to represent the whole document collection [4]. We tried to keep as much information from the original text of every document without modifications as possible. Our method takes into consideration that sentence structure and word order carry just as much important semantic information to a reader as word appearances. Our approach differs from the article routing method because it can be used without any human annotation and specially constructed profiles of expertise [25].

3. Prototype-matching system and its methodology

Our prototype-matching system is a simple content-based IR system. The system aims to retrieve the documents that contain *the same meaning* from the entire document collection. The prototype-matching system analyzes a document collection structure and thus is domain adjusting. The system consists of three parts: document collection preprocessing and encoding, document processing and matching, and document retrieval.

3.1. Document collection preprocessing and encoding

a. Pre-processing takes place before text documents are presented to the text clustering system. We do a basic filtering so that every sentence occupies its own line. Compiling the abbreviation file performs synonym and compound word filtering. We round numbers, separate punctuation marks with spaces, and exclude extra carriage returns, mathematical signs, and dashes. We kept the stop words.

b. After basic filtering of the documents in a text collection, we perform bag-of-word encoding of every document in it. Although this encoding approach is accurate and sustainable for statistical analysis, it is sensitive to capital letters and conjugations. Every word *w*

in a document is transformed into a unique number according to the following formula:

$$y = \sum_{i=0}^{L} k^i \times c_{L-i} \quad (1),$$

where L is the length of the word character string, c_i is the ACSII value of a character within a word w, and k is a constant. We empirically choose k equal 256 since we are using 8-bit ACSII character set. The encoding algorithm produces a unique number for each word disregarding word stems, capitalization and synonyms, so that only the same word can get an equal number. The codes of every word and every single punctuation mark from every document formed feature word vectors representing individual documents stored in the file that corresponds to this document.

3.2. Document processing and matching

We have used the text clustering methodology and vector quantization algorithm for document processing and matching on word and sentence levels [23] that currently consist of the following steps:

a. We look at distribution (a set of word code numbers from 3.1b) for the entire document collection and compute the minimal and maximal parameters (a and b) for the word codes. In the training phase, we divide the range between the minimal and maximal values of words' code numbers into N_w logarithmically equal bins. We normalize the bins' counts according to the quantity of all words in the text. For estimation of the word codes' distribution, we chose the Weibull distribution. The Weibull distribution - one of the most widely used lifetime distributions in reliability engineering[2] - is a versatile distribution that can take on the characteristics of other types of distributions based on the value of the shape parameter. A number of Weibull distributions are calculated with various possible values for a and b using a selected precision. The best fitting Weibull distribution is compared with the code distribution by calculating the Cumulative Distribution Function according to:

$$CDF = 1 - e^{(((-2.6 \times \log(y / y_{\max}))^b) \times a)} \quad (2),$$

where a and b are the parameters of adjusted Weibull distribution. The size of every bin is $1/N_w$.

Hereby, we have created a common word histogram for the entire document collection. The quantization is the best where the words are the most typical to a text (usually 2-5 symbol words - the most widespread length of English words). The distribution and thus quantization of longer words is sparser.

b. Similarly to the word level, we convert every sentence into a number on the sentence level. First, every word in a sentence is changed to a bin number (bn_i) in the same way as we did for words. The whole sentence is considered as a sampled signal. Since the sentences in the text contain different numbers of words, the sentence vector's lengths vary. To get past this fact we apply Discrete Fourier Transformation (DFT) to convert every sentence vector in a collection into input signal. Then we select coefficient B_i (i=1..n) to represent the transformation and the sentence signal. For these coefficients we create quantization like the one on the word level. The range between the minimal and maximal parameters of the sentence code distribution is divided into N_s equally sized bins. We calculate the frequency of sentences belonging to each bin. Then we divide the bins' counts with the total number of sentences in a collection. Finally, we find the best Weibull distribution corresponding to both cumulative distributions. A graphical representation of a sentence quantization process is given in [23].

c. Furthermore, we examine every document in a collection by creating the histograms of the documents' word and sentence code numbers (levels), according to the corresponding values of quantization. We encode the filtered document from a collection word by word on the word level. Each word code number is quantified using word quantization created with all the words in the database. The histogram consists of N_w bins and is normalized by the total number of words in the document. We create similar histograms for every document in the database for the sentence level.

3.3. Document retrieval

Using the histogram of all the documents in the collection, we analyze the single documents' text on the word and sentence levels, in order to compare them using any distance measures. The closest documents in terms of the smallest Euclidian distance between them form a cluster. To complete the retrieval part we choose the documents with the smallest distances to the prototype. The system creates a distance the proximity table of all distances among the documents in a collection. We retrieve the documents from the top of proximity table to every prototype document presented to the system within the set recall window.

4. Description of Task and Prototype Software

One of the distinct features of the Hawaii International Conference on System Science is cross-topic research.

[2] http://www.weibull.com (1998)

The interdisciplinary nature of research creates certain obstacles within decision-making concerning what track a particular paper belongs to.

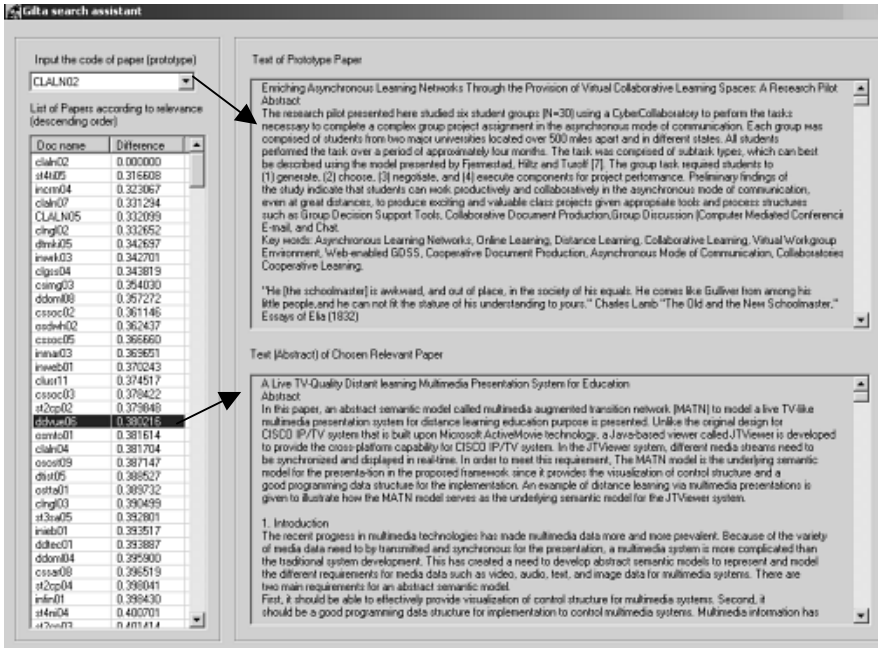

Figure 1. User Interface

The conference organizers have noticed the similarities that run across the traditional tracks in the submitted papers and tried to schedule those papers carefully to avoid conflicts. Trying to save the efforts of processing submitted papers manually, the conference organizers often rely on an authors' presentation of the keywords and headings as a reflection of the main topic of a paper, or on authors' choice of the submission minitrack, or on the track leaders who decide whether a particular paper is relevant to a track stream. All of those approaches risk the occurrence of incorrect paper classification and decrease the conference attendees' satisfaction. We offer our user the opportunity to input into the system the paper (prototype) from a conference collection he/she has an interest in, and thereby, to retrieve the papers that are semantically close to it. The user uses a prototype (a whole paper, or its abstract) instead of spending time on constructing a smart query.

The interface of our running prototype software is depicted in Figure 1. On the left panel we present the conference codes of a prototype paper (e.g. "Enriching Asynchronous Learning Networks Through the Provision of Virtual Collaborative Learning Spaces: A Research Pilot" (CLALN 02)) and the top of its distance proximity table with the codes of closest submitted papers (e.g. "webXice: an Infrastructure for Information Commerce on the WWW" (ST4TI05), "Conceptualizing Trust: A Typology and E-Commerce Customer Relationships Model" (INCRM04), etc.). We present the unfiltered texts of the query (a prototype paper) and a chosen retrieved paper ("A Live TV-Quality Distant learning Multimedia Presentation System for Education" (DDVUE06) that is close to a prototype on the upper and lower right panels respectively.

Our task is similar to an article routing task from [25] with *n* classes and a set of *m* articles. The task attributes each of *m* articles to one of the *n* classes, i.e. each paper has to go to one track/theme. The mathematical tasks of article routing and paper allocation for retrieval by content are similar. However, our interpretation of the task is different in the sense that automatic routines do not seek to establish similarities between *m* classes (tracks/theme) and *n* articles as we do. The differences between *m* classes in article routing task are known or can be learned from a previously constructed profile of expertise. We rely solely on our system's retrieval ability to determine paper content similarities and establish tracks and themes.

5. Description of Data Collection

As an experimental data set for our prototype-matching system, we have chosen all 444 scientific papers obtained from HICSS-34, consisted of about 4440 pages of scientific information, and occupied approximately 14 MB of hard drive space. Looking for the common topics and cross-links in this collection manually would be a time-consuming task. The scientific papers at HICSS-34 were arranged into nine major tracks, which were further divided into seventy-eight minitracks. The organizers made an effort to identify six themes that run across the traditional tracks based on the similarities and expansion of the scientific fields. Table 1 presents the taxonomy of the HICSS-34 conference. The outlined six *cross-track themes* are listed at the bottom of Table 1. They covered 168 papers in the conference from 30 mini-tracks. An organizing committee assigned a unique identification code to every paper. The code shows what track and minitrack the particular paper belongs to. We processed papers in portable-document format, which were converted to plain ASCII text. Distinct regions of the papers (title, authors, abstract, main body and bibliography) were manually identified and extracted, so

only title, abstract and main body were left remaining in a filtered document collection for further experiments.

№	Track Title /Number of papers /Number of Minitracks
1	Collaboration Systems and Technology /66 /9
2	Complex Systems /29 /5
3	Decision Technologies for Management /47 /7
4	Digital Documents /40 /6
5	Emerging Technology /30 /4
6	Information Technology in Health Care /26 /5
7	Internet and Digital Economy /68 /12
8	Organizational Systems and Technology /63 /14
9	Software Technology /75 /13
№	**Theme Title /Number of papers in it**
1	Knowledge Management/20
2	Data Warehousing-Data Mining/24
3	Collaborative Learning/22
4	Workflow/12
5	E-commerce Development/54
6	E-commerce Application/36

Table 1. HICSS-34 Tracks and Themes Taxonomy

6. Experiments

We have conducted several separate experiments to test the prototype-matching system capability to allocate the scientific conference papers based on their content. In our experiments, we have used the methodology described in Section 3 and the HICSS text collection described above. We have examined the ability of the prototype-matching system to retrieve the most similar papers to a presented prototype from the entire conference collection. We have expected to have papers either from the same track or from the same cross-track theme among the closest matches to a certain prototype. We have tried different sizes of recall window. We have not considered the order within a window, only paper co-occurrence, using precision and recall as effectiveness measures [22].

6.1 Experimental Settings

In the first experiment we have examined the consistency of the tracks proposed by conference organizers using every paper from every track as a prototype and a query. We have evaluated how often papers from the same track fire as the closest matches to the prototype-paper from the same track. We have analyzed tracks one by one, making a comparison between our prototype-matching clustering and the track division, proposed by the conference organizers.

The second experiment had a slightly different scope. We have examined the consistency of the cross-track themes, proposed by conference organizing committee as the semantic subsets of the papers from different track, with different key words and lexis. The division was meant to unite the papers with different terminologies and headings from different tracks into one interdisciplinary research theme. We have used every paper in every theme as a prototype and a query, trying to retrieve the most semantically similar papers. We have analyzed the cross-track themes one by one, comparing our prototype-matching clustering with the conference theme division. Because we built our system so that it could detect not only word co-occurrence but also main semantic similarities, we anticipate the second experiment give clustering results close to the conference theme division. We have expected to have the closest matches from different tracks but from the same cross-track theme.

6.2 Results

After presenting every paper as a prototype to the system we have obtained the proximity tables. The proximity table is a matrix of distances between a prototype and the rest of the papers in a collection. We present the fragment of a proximity table and the line of reasoning about some results below. Table 2 contains an example of a proximity table (recall window 23) for "Enriching Asynchronous Learning Networks Through the Provision of Virtual Collaborative Learning Spaces: A Research Pilot" (CLALN02), "Studies of ALN: An Empirical Assessment" (CLALN05), "CTER OnLine: Evaluation of an Online Master of Education Focusing on Curriculum, Technology and Education Reform" (CLALN06), and "A comparative Content Analysis of Face-to-Face vs. ALN-Mediated Teamwork" (CLALN07). All those papers belong to the Collaboration Systems and Technology Track, the Asynchronous Learning Networks Minitrack and, additionally, are assigned to Collaborative Learning Theme. The taxonomy of this theme is presented in Table 3. In Table 2 we highlighted the codes of papers that belong to the same cross-track theme as our sample papers using gray background. We used the italic font to outline the codes of papers from the minitracks of the tracks that form the Collaborative Learning cross track theme. Paper CLGSS04 is underlined for the reasons stated in the discussion section.

One can notice that the paper CLALN02 has CLALN05 and CLALN07 among its closest matches, CLALN05 has CLALN02 among its closest matches, but at the same time, CLALN07 has none of those papers among its closest matches. The distances between CLALN07 and its closest matches in comparison with distances between CLALN02 or CLALN05 and their closest matches respectively, show that CLALN07 has

papers other than CLALN02 and CLALN05 that are semantically closer. The distance range for CLALN 02 is [0.317..0.387], but at the same time, the distance range for CLALN07 is [0.249..0.329] (recall window=23). By the implication logic, the paper CLGSS04's closeness to CLALN02, CLALN05, and CLALN07 proves that all those papers are semantically similar.

Codes of the papers that are closest matches to a prototype	Codes of prototype-papers							
	CLALN02	0	**CLALN05**	0	**CLALN06**	0	**CLALN07**	0
	ST4TI05	0.316	**ST3SE03**	0.265	**ST4NI04**	0.304	***CLGSS04***	0.248
	INCRM04	0.323	**ST4TI05**	0.322	***CLDGS07***	0.334	***DDOML08***	0.254
	CLALN07	0.331	***DDPTC09***	0.328	**OSTTA05**	0.337	***DTMKI03***	0.270
	CLALN05	0.332	**CLALN02**	0.33	***DDOML04***	0.338	**ST2CP07**	0.271
	CLNGL02	0.332	***DDOML11***	0.334	**OSRMA02**	0.339	**CSSOC03**	0.279
	DTMKI05	0.342	**INWEB01**	0.334	**ST2IM01**	0.339	**OSTOI02**	0.282
	INWRK03	0.343	**OSMTO01**	0.341	***DTMKI03***	0.340	**INCRM01**	0.292
	CLGSS04	0.344	***CLUSR23***	0.346	**OSSCI03**	0.347	**ST2CP04**	0.295
	CSIMG03	0.354	***CLUSR14***	0.347	**ST2CP04**	0.348	***DTMKI05***	0.296
	DDOML08	0.357	**ST2CP03**	0.358	**ETEGV05**	0.355	**ST3SE01**	0.297
	CSSOC02	0.361	**CSSOC03**	0.359	**INMAR05**	0.356	***DTUML04***	0.306
	OSDWH02	0.362	**DDVUE05**	0.364	***DDOML08***	0.359	***DDUAC03***	0.307
	CSSOC05	0.366	**INWRK03**	0.365	**OSINF04**	0.360	***CLUSR17***	0.309
	INMAR03	0.369	**CLNGL02**	0.366	***DTUML03***	0.361	**CSSOC02**	0.311
	INWEB01	0.370	***DTMKI05***	0.368	**INWEB06**	0.362	**INWRK03**	0.312
	CLUSR11	0.374	***DTMKI03***	0.369	**CLALN07**	0.364	**OSPMT05**	0.313
	CSSOC03	0.378	***CLCDV06***	0.370	**INCRM04**	0.367	**ST2CP01**	0.315
	ST2CP02	0.379	**CSSAR08**	0.373	***DTUML04***	0.368	**INCDE05**	0.317
	DDVUE06	0.380	**ST3SA05**	0.373	***DTMKI05***	0.370	**OSCIS01**	0.318
	OSMTO01	0.381	***CLGSS04***	0.376	**OSETH01**	0.371	**OSPMT01**	0.323
	CLALN04	0.382	***DDOML08***	0.377	**INCRM01**	0.375	**ST1MA05**	0.326
	OSOST09	0.387	**CLNGL03**	0.378	**INWEB05**	0.381	**INWEB01**	0.328

Table 2. A Fragment from a Proximity Table for papers CLALN02, CLALN05, CLALN06, CLALN 07 from "Collaborative Learning Theme" (Recall window =23)

Name of the Track	Name of the Minitrack within a Theme (code of papers in it)
Collaboration Systems and Technology Track	Next Generation of Learning Platforms (CLNGL01-03)
	Asynchronous Learning Networks (CLALN01-07)
	Technology Supported Learning (CLTSL01-03)
Digital Documents Track	Digital technology and Educational Culture (DDTEC01-03)
	Digital Documents in the Office and Education (DDVUE01-06)

Table 3 Taxonomy of Collaborative Learning Theme

6.2.1. The First ("Track") experiment. In the first experiment, we have detected all closest matches to every paper from the conference collection. Aiming to check the consistency of conference tracks, we have focused on the results retrieved by our systems for every of nine tracks. We have presented every paper from the collection as a prototype to the system and calculated the "hit ratio", which reflects how often a paper from the same track has fired as the closest match to the presented prototype in a given recall window. We assumed that the tracks should be somewhat balanced by the number of papers in them, and used a recall window of 47, as the average size of HICSS tracks and at 25 for reasons stated later. Table 4 contains hit ratios per track (hit ratio1 and hit ratio2), that reflect how many papers from the same track were retrieved among the 47 or 25 closest matches respectively.

Track Title	Number of Papers	Max Hit ratio1	Max Hit ratio 2
Collaboration Systems and Technology	66	22.7%	15.2%
Complex Systems	29	17.2%	13.8%
Decision Technologies for Management	47	19.1%	12.8%
Digital Documents	40	22.5%	15%
Emerging Technology	30	20%	13.3%
Information Technology in Health Care	26	23.1%	15.4%
Internet and Digital Economy	68	19.1%	13.2%
Organizational Systems and Technology	63	22.2%	15.9%
Software Technology	75	22.7%	14.7%

Table 4. The results from track division clustering (Recall window = 47 and 25)

6.2.2 The Second ("Theme") experiment. In the second experiment we have focused on the closest matches to 168 papers that were arranged by the conference organizers into six cross-track themes. We expected to have many closest matches from the same cross-track theme in the recall window, because themes are meant to unite semantically close papers. Assuming that the conference committee wishes to have roughly balanced themes, we set a recall window at 25. Table 5 contains the name and sizes of cross-track themes, and hit ratios that reflect how many papers from the same theme fired among the 25 closest ones to a prototype paper from the same theme. The last column (hit ratio4) shows how many papers have their closest matches from the same track, some minitracks of which have formed the certain theme. For instance, the "Collaborative learning" theme has the highest hit ratio at 31.8%. This means that almost every third paper among the closest matches was from the same theme as a prototype paper. The Collaborative Learning theme has the highest hit ratio4 at 23.2%, stating that almost every fourth paper among the 25 closest matches was from Collaboration Systems and Technology or

Digital Documents tracks, maybe from minitracks different than mentioned in Table 3.

Theme Title	Number of Papers	Max Hit ratio3	Max Hit ratio4
Knowledge Management	20	25%	8.46%
Data Warehousing/ Data Mining	24	20.8%	9.7%
Collaborative Learning	22	31.8%	23.2%
Workflow	12	18.2%	10.5%
E-commerce Development	54	18.5%	11.4%
E-commerce Application	36	17.1%	8.6%

Table 5. The results from cross-topic theme clustering (Recall window = 25)

6.3. Discussions

We set the same size of recall windows to make the results from both experiments comparable. Looking at hit ratio2 from Table 4 and hit ratio3 from Table 5 we conclude that the hit ratios for theme division, on average, were slightly higher than hit ratios for track division with the same size recall window. This demonstrates the stronger semantic similarity among the papers from the same cross-track themes than the semantic similarities among the papers from the same tracks. However, the hit ratio for "E-commerce Development" and "Workflow" themes are rather low in comparison with hit ratios of the other themes. After analyzing those themes we discovered that some papers in them contain the initial data that was treated as noise by our system. Although the paper "Workflow Analysis Using Attributed Metagraphs" (ST2IM05) has fired in the bottom of a proximity table to all 11 papers from the same theme, it clearly belongs to the Workflow theme, because it discusses presentation and formal analysis of workflows as metagraphs with specified temporal constraints for time-critical tasks, i.e. for generalization of traditional network scheduling methods used in project management. The paper is full of technical details and formulas that our system, apparently, failed to understand. The "E-commerce Development" theme has several outliers that do belong to the theme semantically, but use very diverse lexicon (e.g. on-line markets instead of electronic or e-market), such as "Second-Degree Price Discrimination for Information Goods Under Nonlinear Utility Functions" (INEEC06) and "Transforming Financial Markets to Retail Investors: A Comparison of the U.S. and the German On-line Brokerage Market" (INFIN02). The paper INEEC06 contains a mathematical description of a model for price discrimination followed by heavy mathematical reasoning from one formula to another that our system fails to recognize. The paper INFIN02 has fired at the bottom of the proximity table to 16 papers from the same cross-track theme because it uses many unique proper adjectives, such as German, European and American that influences the sentence construction. The change in sentence construction has an impact on the retrieval ability of our system. In order to enhance our IR system's abilities to handle synonymous attributes and disambiguation we plan to construct an extensive synonym table for filtering.

We have noticed that word usage and some peculiarities of the academic written style of the scientific papers in a collection have a significant impact on the clustering ability of our method. All research papers consist of the same components: introduction, method, research background, results and discussion [7]. Therefore the intervals of distance measures on word and sentence level are so narrow ([0.314...0.773] and [0.23...1.149] respectively). A particular academic writing style, since authors tend to use similar word order and similar sentence structures to describe their achievements in information system research, e.g. *we present, computer analysis our paper discusses, construct a model, approach is based, process information, in the remainder of a paper, this paper describes, traditional systems, etc*. explains the closeness of all papers on the sentence level. Finally, we have discovered that our prototype-matching clustering of the scientific text corpus is somewhat different from the theme division proposed by the organizing committee. It can be explained either by poor allocation of papers to minitracks, tracks or themes by the HICSS organizers, or by poor performance of the proposed method.

One can notice that the good matching results of automated clustering with human's manual selection should be in the range [40%…60%] versus our experimental results with matching range [13%…32%]. The justification of the ranges is not an obvious task since for the calculations of matching results we compared our retrieval results to the track and theme divisions provided to us by the conference organizing committee. Those divisions can be a product of the message that every paper conveys, the author's vision of a paper and a number of non-optimal considerations that conference chairs keep in mind in addition to the topic relevance of a paper submitted to a certain track or theme. As one subcommittee chair has noticed, there are a number of other issues in addition to content relevancy that should be balanced in conference settings, such variables as conflict of interests, gender, geography, topic, etc. Yarowsky and Florian noticed that members of conference committees tend to favor the article with the most interesting content and findings and route them to their tracks, even if the topics are not so relevant [25]. Obviously, our prototype-

matching system that is based on objective text processing technique does not consider those issues. The nature of hit ratios' calculation makes the evaluation of out results very challenging. We calculated the hit ratios on the strong assumption that a given theme/track division by the HICSS committee is the semantically absolutely correct one. However, the HICSS theme/track division is a very weak reference point for comparison because of the issues mentioned above. For instance, CLALN02 and DDVUE06 papers that belong to the same cross-track theme "Collaborative Learning", apparently convey different messages (see Figure 1).

Additionally, there are a number of useful subtasks that our prototype-matching system can handle. It can be used to detect "good-candidate" papers to be included in the theme from minitracks that were not included in it. For example, the paper "The Mindpool Hybrid: Theorizing a New Angle on EBS and Suggestion Systems" (CLGSS04) from the Group Support Systems minitrack has fired at the top of the proximity table to almost every third paper from Collaborative Learning theme (shown in Table 2). Careful reading of this paper has shown that it could be included in the theme because it discusses the improvements possible that can be achieved by using computer support in electronic brainstorming, which is a collaborative technique in nature.

Our system can have another use, similar to article routing discussed in [25]. The prototype-matching system can mark submitted papers and send them to the appropriate conference subcommittee or minitrack. Then we need to add into our document collection the descriptions of minitracks that are usually provided by minitrack chairs. There are fewer possibilities for conflicts of interest involving the balance between the issues discussed above. The users of the system could use this method in a semi-automatic setting where the program makes recommendations for paper rerouting but it is up to the conference or track chairs to make the final decision.

As for the limitations of our study, we can consider the critique toward the scalability of the methodology, small experimental data collection and result evaluation. However, the methodology evaluation was offered in [23] by examining the similarities in different translation of the books of Bible. The question is how well does the new proposed system compare to the standard (manual) methods of doing this, both in terms of costs (manual work by the chairs, subcommittee chairs, authors, etc) and benefits has to be explored.

7. Conclusions and Future Work

In this paper we have clustered a scientific text collection from the Hawaii International Conference on System Science-34 using the prototype-matching IR system. The conference organizers of HICSS-34 had offered non-traditional cross-track theme classification of the submitted papers to help the conference attendees to visit all sessions relevant to their research needs. We aimed to allocate semantically close papers from HICSS according the non-traditional conference taxonomy. Our prototype-matching IR system consists of document encoding, prototype matching and retrieval parts. The core of prototype-matching system of text filtering, "smart" document encoding on word and sentence levels, creating word and sentence level histograms, and prototype matching phases. We form clusters according to the Euclidian distances between the prototype article and the rest of a document collection.

In the paper we have presented two experiments from the clustering sessions on a scientific collection. We tested the system's ability to retrieve the closest papers according to content from the whole document collection. In our first experiment we examined track consistency with respect to the retrieved results from our system. In the second experiment, we examined the semantic closeness of the papers within every cross-track theme by studying their retrieved closest matches. Even though our clustering results turned out to be somewhat different from the cross-track division offered by the conference organizers, our method was able to capture some semantic similarities between the scientific papers. The specific limited vocabulary and conservative academic style of the scientific papers had a strong impact on our clustering results.

We suggest the use of our system's prototype-matching clustering ability for processing a big number of text documents during the limited period of time. Reading some of the chosen papers in each cluster can provide the decision maker with the main ideas of all the documents from this cluster. As future work, we consider evaluation of the system's effectiveness with the help of expert-readers, methods for improvements, and exploration of additional uses of the system. In the long-term prospective, the system can be used as an on-line retrieval system that can help conference participants to choose the most interesting conference venues.

8. Acknowledgement

We gratefully acknowledge the financial support of TEKES (grant number 47 533) and the Academy of Finland.

9. Reference

[1]. Anderson, N., "A Tool for Building Digital Libraries", *Digital Library Journal Review* 5, (2), 1999

[2]. Anick, P., Vaithyanathan, S., "Exploiting Clustering and Phrases for Context-Based Information_Retrieval" SIGIR 97, Philadelphia, USA, ACM,.1997

[3]. Chien, L. F., "PAT-Tree-Based Keyword Extraction for Chinese Information Retrieval", Proceedings of Special Interest Group on Information Retrieval, SIGIR'97, Philadelphia, USA, ACM Press, 1997.

[4]. Cutting, D., Karger, D., Pedersen, J., and Turkey, J., "Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections", 15th Annual International SIGIR'92, Denmark, ACM Press, NY, USA, 1992

[5]. Deogun, J., Raghavan, V., User-oriented document clustering: a framework for learning in information retrieval, ACM conference on Research and development in information retrieval, Pisa, Italy, ACM Press New York, NY, USA, 1986

[6]. Dewey, M., *A classification and subject index for cataloguing and arranging the books and pamphlets of a library,* Amherst, MA, USA, Case, Lockwood & Brainard Co., 1876

[7]. dos Santos, M., "The textual organization of research paper abstracts in applied linguistics", *Text* 16(4), 1996, pp 481-499

[8]. Hand D., M. H., and Smyth P., *Principles of Data Mining. Boston*, USA, A Bradford Book, The MIT Press, 2001.

[9]. Hatzivassiloglou, V., Gravano, L., Maganti, A., "An Investigation of Linguistic Features and Clustering Algorithms for Topical Document Clustering", The 23rd ACM/SIGIR conference on Research and development in IR, Athens, Greece, ACM Press New York, USA, 2000

[10]. Jo, T., "Text Categorization considering Categorical Weights and Substantial Weights of Informative Keywords", Tokyo, Japan, Samsung SDS, 1999, pp1-17

[11]. Kohonen, T., "Self-Organization of Very Large Document Collections: State of the Art", Proceedings of ICANN98, the 8th International Conference on Artificial Neural Networks, Springer, London, 1998

[12]. Kohonen, T., *WEBSOM*. Helsinki, Helsinki Technological University, Finland, 1999

[13]. Lawrence, S., Bollacker, K., Lee Giles, C., "Indexing and Retrieval of Scientific Literature", The 8th International Conference on Information and Knowledge Management (CIKM 99), Kansas City, MO, USA, ACM Press, 1999

[14]. Lee, C., and Yang, H., "A Web Text Mining Approach Based on Self- Organizing Map", WIDM-99, Kansas City, MO, USA, ACM Press, 1999

[15]. Lin, X., Soergel, D., and Marchionini, G., "A Self-organizing Semantic Map for Information Retrieval", The 14th ACM/SIGIR conference on Research and development in IR, Chicago, IL, USA, ACM Press, 1991.

[16]. Merkl, D., and Schweighofer, "En Route to Data Mining in Legal Text Corpora: Clustering Neural Computation, and International Treaties", The 8th International Workshop on Database and Expert Systems Applications (DEXA'97), Toulouse, France, IEEE, 1997

[17]. Miike, S., Etsuo, I., Ono, K., Sumita, K., "A full-text retrieval system with a dynamic abstract generation function", The 17th ACM/SIGIR conference on Research and development in IR, Dublin, Ireland, Springer-Verlag New York, Inc., 1994

[18]. Salton. G., a. M. McGill (1983). *Introduction to modern information retrieval.* New York, McGraw-Hill.

[19]. Schutze, H., and Silverstein, C., "Projection for Efficient Document Clustering", ACM/SIGIR-97, Philadelphia, PA, USA, ACM Press New York, USA, 1997

[20]. Sparck-Jones, K., *Automatic Keyword Classification for Information retrieval,* Connecticut, Archon Books, 1971

[21]. Sprague, R. H., Jr., "Preface to The Hawaii International Conference on System Science 2001", HICSS-34, Maui, Hawaii, 2001

[22]. van Rijsbergen, C., *Information Retrieval* (Second Edition), London: Butterworths, 1979

[23]. Visa, A., Toivonen, J., Back, B., and Vanharanta, H., "Contents Matching Defined by Prototypes: Methodology Verification with Books of the Bible." *Journal of Management Information Systems* 18(4), 2002, pp 87-100

[24]. Witten, I., Bray Z., Mahoui, M., and Teahan, B., "Text mining: A new frontier for lossless compression",. Data Compression Conference '98, IEEE, 1998

[25]. Yarowsky, D. and R. Florian, "Taking the load off the conference chairs: towards a digital paper-routing assistant", Joint SIGDAT Conference on Empirical Methods in NLP and Very Large Corpora, ACM Press, 1999

0-7695-1874-5/03 $17.00 © 2002 IEEE