

REPRESENTING INFORMATION COLLECTIONS FOR
VISUAL COGNITION

A Dissertation

by

EUNYEE KOH

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

August 2008

Major Subject: Computer Science

REPRESENTING INFORMATION COLLECTIONS FOR
VISUAL COGNITION

A Dissertation

by

EUNYEE KOH

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee,	Andruid Kerne
Committee Members,	Richard Furuta
	Ricardo Gutierrez-Osuna
	Yu Ding
Head of Department,	Valerie E. Taylor

August 2008

Major Subject: Computer Science

ABSTRACT

Representing Information Collections for

Visual Cognition. (August 2008)

Eunyeek Koh, B.S., Seoul National University

Chair of Advisory Committee: Dr. Andruid Kerne

The importance of digital information collections is growing. Collections are typically represented with text-only, in a linear list format, which turns out to be a weak representation for cognition. We learned this from empirical research in cognitive psychology, and by conducting a study to develop an understanding of current practices and resulting breakdowns in human experiences of building and utilizing collections. Because of limited human attention and memory, participants had trouble finding specific elements in their collections, resulting in low levels of collection utilization. To address these issues, this research develops new collection representations for visual cognition. First, we present the *image+text surrogate*, a concise representation for a document, or portion thereof, which is easy to understand and think about. An information extraction algorithm is developed to automatically transform a document into a small set of image+text surrogates. After refinement, the average accuracy performance of the algorithm was 90%. Then, we introduce the composition space to represent collections, which helps people connect elements visually in a spatial format. To ensure diverse information from multiple sources to be presented evenly in the composition space, we developed a new control structure, the *ResultDistributor*. A user study has demonstrated that the participants were able to browse more diverse information using the ResultDistributor-enhanced composition space. Participants also found it easier and more entertaining to browse information in this representation. This research is applicable to represent the information resources in

contexts such as search engines or digital libraries. The better representation will enhance the cognitive efficacy and enjoyment of people's everyday tasks of information searching, browsing, collecting, and discovering.

To My Family

ACKNOWLEDGMENTS

I welcome this opportunity to thank the many people who supported me to the completion of my Ph.D. project and dissertation. First, I want to thank my advisor. I cannot imagine myself completing this research without my advisor. He has been an endless supporter, teacher, and role model. I also want to thank my committee members who have provided important feedback that has helped me to complete this research.

My family, essential in my life, has been providing me with all the necessary things that graduate students normally cannot afford and have also offered mental support that helped me to get through all the hardships during these five PhD years. Many friends have been working with me to conduct projects, to develop ideas, and to discuss about PhD lives as well. Having nice, creative and smart friends in my school year was very fortunate for me. They sometimes challenged me, helped me, and laughed with me, cheering great successes that we accomplished together. They are too many to list here, but I would like to say thank you to all of them, especially to friends in the Interface Ecology Lab.

TABLE OF CONTENTS

CHAPTER		Page
I	INTRODUCTION	1
II	BACKGROUND	7
	A. Surrogate	7
	B. Surrogate Representations	8
	C. Cognition of Images and Text	9
	D. Composition Space	10
	E. combinFormation: Mixed-Initiative Composition	10
III	I KEEP COLLECTING	14
	A. Background	16
	B. Study Description	17
	C. Study Results	20
	1. Collection Building and Utilizing	21
	2. Intention and Need	21
	3. Activities and Significance	21
	4. Frequency and Time Period	21
	5. Worthwhile or Useless	22
	6. Collection Types	22
	7. Collection Mechanism	23
	8. Levels of Engagement with Collections	24
	9. Collection Sharing	26
	10. Breakdowns in Collection Practice	26
	11. Reasons for Collection Building	28
	12. Using Semantics to Represent Collections	28
	13. Developing Informal Metadata Schemas	29
	14. Suggestions	30
	D. Implications for Design	30
IV	INFORMATION EXTRACTION ALGORITHM: SURRO- GATE CLASSIFICATION THROUGH PATTERN RECOG- NITION	35
	A. Background	35

CHAPTER	Page
B. Surrogate Features	38
C. Document Surrogate Model	40
D. Pattern Recognition Approach	43
1. Pattern Classifier	44
2. Cross-Validation Method	45
E. Experiments	45
1. Datasets	45
2. Results	47
3. The Structured Collection	47
4. The Non-Structured Collection	50
5. The Complete Set	50
F. Discussion	55
V TEST COLLECTION MANAGEMENT AND LABELING SYSTEM	57
A. Background	58
B. System Design	60
1. Document Labeling Semantics	60
2. Interactive Collecting and Labeling Client	63
3. Identify Each Document to Collect	63
4. Store the Document in Repository	66
5. Label Each Document	68
6. Store Labels in Repository	69
C. Building the Test Collection	70
D. Discussion	71
VI INFORMATION EXTRACTION ALGORITHM: EXTRACT- ING IMAGE+TEXT SURROGATES VIA RECOGNIZING INFORMATIVE CONTENT FROM WEB PAGES	73
A. Background	76
B. Information Extraction Algorithm	76
1. Categorize Index Page or Content Page	77
2. Recognize Informative Contents	79
3. Extract Representative Images and Text	81
C. Experiments	82
1. Dataset	83
2. Evaluation Metrics	83
3. Experimental Results: Page Categorization	84

CHAPTER	Page
4. Experimental Results: Informative Content Body Detection	87
5. Experiment Results: Informative Image Detection	89
D. Discussion	89
VII RESULTDISTRIBUTOR: GENERATING DIVERSE INFORMATION IN MIXED-INITIATIVE COMPOSITION	92
A. ResultDistributor	93
1. Yahoo Buzz with ResultDistributor	99
B. Balancing the Search Processing with Results	99
C. ResultDistributor Evaluation with Mixed-Initiative Composition	101
1. Study Apparatus	102
2. Participants	103
3. Procedure	104
4. Study Design	105
5. Results	105
D. Background	110
E. Discussion	111
VIII CONCLUSION	112
A. Image+Text Surrogate Extraction	113
B. Implications for Search Engines	115
C. Further Improvements: Recognize Index-Content Pages	117
D. Increasing Diversity in Mixed-Initiative Composition Space	120
E. Summation	121
REFERENCES	123
VITA	132

LIST OF TABLES

TABLE		Page
I	Document surrogate features.	40
II	Test data characteristics and performance results with cross validation.	50
III	Document labeling semantics for the test collection to validate informative images and text extraction algorithm.	61
IV	Contingency table with the news collection: the experimental results of the page categorization algorithm. ‘Correct’ means the correct category of the pages, and ‘Assigned’ means the category assigned by the algorithm.	84
V	Performance of page categorization algorithm with the news collection.	84
VI	Contingency table with the research collection: the experimental results of the page categorization algorithm. ‘Correct’ means the correct category of the pages, and ‘Assigned’ means the category assigned by the algorithm.	85
VII	Performance of page categorization algorithm with the research collection.	85
VIII	Contingency table with total test collection: the experimental results of the page categorization algorithm. ‘Correct’ means the correct category of the pages, and ‘Assigned’ means the category assigned by the algorithm.	86
IX	Performance of page categorization algorithm with total test collection.	86

LIST OF FIGURES

FIGURE	Page
1	Top: A search result snippet from Google; Bottom: An enhanced search result snippet with an informative image extracted from the Wikipedia page. This is an example of an image+text surrogate. 2
2	Example of a composition space created with the mixed-initiative system, combinFormation by a student in the ENDS course. 3
3	Concept map: organization chart of the dissertation structure. 5
4	Linear list of Google surrogate: A typical surrogate representation. 8
5	A composition of image and text surrogates representing sets of information resources for an overview of part of the undergraduate psychology curriculum. Each surrogate is formed by clipping information elements from source containers. 11
6	Mixed-initiative composition: Interactions of human beings and procedural generation of digital information in the mixed-initiatives visual composition space. 12
7	College students' digital collection examples. Top-Left: a personal paper collection saved in file folders. Top-Right: a published personal collection of papers on the web to share with others. Bottom-Left: a personal image collection in a file folder. Bottom-Right: a published personal collection of images on the web for sharing. 18
8	Number of elements participants collect, by media type. 23
9	Top - participants' Internet access and collection building, referring, and organizing frequency; Bottom - rate at which participants' collections are unutilized and abandoned. 25
10	2D PCA scatter plot of text surrogate candidates in the Structured Collection. 48

FIGURE	Page
11	2D PCA scatter plot of image surrogate candidates in the Structured Collection. 49
12	2D PCA scatter plot of text surrogate candidates in the Non-Structured Collection. 51
13	2D PCA scatter plot of image surrogate candidates. 52
14	2D PCA scatter plot of text surrogate candidates in the Complete Set. 53
15	2D PCA scatter plot of image surrogate candidates in the Complete Set. 54
16	Example of nodes highlighted with Modified DOM Inspector and labels assigned to the test collection document. 62
17	Modified DOM Inspector: URL to Server button stores document in repository. Semantic buttons (right) label selected HTML element. Save XML button stores labeling in repository. 64
18	Above: an example test web page (index.html) and associated resource image (_42687225_matty_pa203b.jpg) stored in the repository; Below: the directory structure stored for the Above test web page in the repository. 65
19	Generate <code>tag_id</code> of each HTML tag by traversing the DOM tree with DFS algorithm. The left diagram shows the DOM tree, with each <code>tag_id</code> generated by algorithm (right). 67
20	Interface for labeling the informative image with its caption in the Modified DOM Inspector. 68
21	Three stages of our information extraction algorithm. Stage 1 determines the page categorization, Stage 2 recognizes the informative sub-tree of the content body page, and Stage 3 extracts representative images and text from the sub-tree. 74
22	Top: Index page; Bottom: Content page with most informative content body selected. 75

FIGURE	Page
23	The procedure of how the algorithm recognizes the content body node by identifying the highest rank nodes in the DOM. 82
24	One way to resolve this failure is to extract informative elements from the pages because the failed pages do contain the representative images and text, or we will be able to refine our algorithm and metrics to determine these pages as index pages. One possibility is that after the algorithm determines the content body, it ranks the content body node to determine whether it is the true content page or not by checking the link threshold. 88
25	An example page that failed in determining the content body node. The outer rectangle border shows what is labeled as the content body block, and the inner rectangle border shows what the algorithm determined as the content body. 90
26	Google’s OR search results from a query, “apple OR orange”. It only retrieves two results about “orange” among the first ten search results. 95
27	Three concurrent processes in the ResultDistributor: (1) Process Search: add results into the appropriate ResultSlice; (2) Process Result: download and extract media from results in the ResultSlice. Move to the next ResultSlice after it finishes all results in the current slice; (3) Adjust the expected number of results in each ResultSlice. 96
28	Pseudo code for the ResultDistributor. 97
29	Browsing TV Leaders in Yahoo Buzz in combinFormation. Large texts in rectangle boxes are labels added to identify the search queries of underlying clustered media elements. 100
30	Mixed-Initiative Composition enhances the diversity of media; Left: number of diverse URLs they browsed; Right: Participants’ experience ratings about media diversity. 106
31	Mixed-initiative composition is easier to use, more liked, finds more relevant and interesting media, and is more entertaining to experience than the typical toolset. 108

FIGURE	Page
32 A flow of how our information extraction algorithm can be used with a typical search engine to enhance the representation of results, and the overall user experience.	116
33 An example of an index-content page which could be recognized as a set of surrogates. Each surrogate consists of a title within a hyperlink, a thumbnail image, and a descriptive paragraph of text. . .	118

CHAPTER I

INTRODUCTION

Due to the increasing popularity of digital media devices and the abundance of information on the web, a broad cross-section of society becomes more and more exposed to large numbers of digital documents and media elements. People collect meaningful digital information, and their collections become larger and larger. This trend is further promulgated by the increasing availability and capacity of inexpensive digital storage devices. However, the representations of collections have not improved accordingly to support these changes. We are working ethnographically, conceptually, and algorithmically to understand people's needs, and develop solutions.

We conducted a user study to understand people's practices in building and utilizing collections, paying particular attention to breakdowns. The study results showed that participants are confronted with the problem of how to keep track of significant elements within the collections, and even though they built collections of elements that were useful, most of them are not utilized in relevant contexts of need because of limits in human attention and memory. To address problems that occurred in collecting practices, this research develops new representations of documents and collections for visual cognition.

We first present the image+text surrogates to represent documents (see Figure 1). Many researchers have demonstrated that the integration of informative image into a typical text-only representation makes better use of human cognitive resources and improve browsing and navigation [1]. Image+text surrogates are formed by transforming each document into a small set of concise representations that are easy to

The journal model is *IEEE Transactions on Automatic Control*.

[Virtual reality - Wikipedia, the free encyclopedia](#)

Virtual reality (VR) is a technology, which allows a user to interact with a computer-simulated environment, be it a real or imagined one. ...

en.wikipedia.org/wiki/Virtual_reality - 89k - [Cached](#) - [Similar pages](#) - [Note this](#)



[Virtual reality - Wikipedia, the free encyclopedia](#)

Virtual reality (VR) is a technology, which allows a user to interact with a computer-simulated environment, be it a real or imagined one. ...

en.wikipedia.org/wiki/Virtual_reality - 89k - [Cached](#) - [Similar pages](#) - [Note thi](#)

Fig. 1. Top: A search result snippet from Google; Bottom: An enhanced search result snippet with an informative image extracted from the Wikipedia page. This is an example of an image+text surrogate.

understand and think about. To automatically form the image+text surrogates, we first developed a *surrogate classification algorithm* to determine whether surrogate candidates, clippings from documents, are informative or not. We developed discriminatory features such as tag patterns with our new *Document Surrogate Model* (DSM) which utilizes only structural elements in the *Document Object Model* (DOM) [2]. The DSM is similar to the “simple, general DOM” (SGDOM) of Phelps and Willensky [3], except that tags involving hyperlinks are considered significant in the DSM, while they are not in the SGDOM. The tag pattern is formed to represent the context of a surrogate candidate by using the immediate tag that defines the candidate’s deepest element in the tree, and a short sequence of ancestor tags. We developed training data and classified informative and non-informative surrogate candidates in a supervised manner using a quadratic classifier [4].

The average classification performance of the surrogate candidate classification algorithm was 80% accuracy. However, we found the three limitations of this algorithm. First, the performance of the surrogate classification was adversely affected by outliers in the training data. Second, the tag patterns do not represent the re-

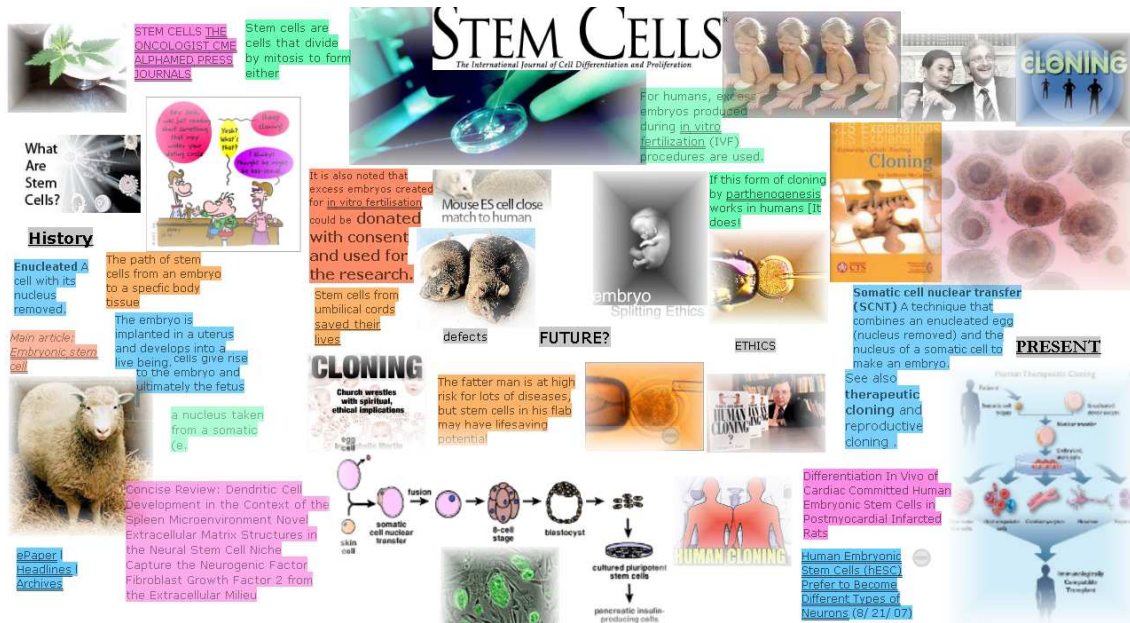


Fig. 2. Example of a composition space created with the mixed-initiative system, combinedFormation by a student in the ENDS course.

relationships (e.g., parent-child, sibling, ...) among different surrogate candidates because the pattern does not uniquely identify a document location. Lastly, the algorithm forms surrogate candidates from all sub-trees in the DOM, and filters the non-informative surrogates later. This produces incorrect results and also is inefficient.

Hence, we proceeded to develop an information extraction algorithm that automatically determines the informative content of a given document by ranking the sub-tree, and then uses the highest ranked sub-tree to extract image+text surrogates. The informative extraction algorithm maintains a compact representation of the tree structure while recognizing informative elements. The algorithm is developed in a decision tree manner that utilizes the document structure, eliminating the dependency on training data by developing decision rules specific to this problem.

The primary decision rule for determining informative content is based on ranking

the nodes of the DOM tree. The ranking metrics are developed with semantic features that discriminate informative content. By finding the common parent node that holds the highest ranked nodes in the DOM tree, the algorithm determines where the informative content are and enables formation of meaningful image+text surrogates.

A labeled test collection for evaluating this algorithm was not previously available. Thus, in order to evaluate its performance, we developed a test collection system, which enables systematically collecting and labeling test data. Using this system, we collected and labeled over 500 web pages to utilize in the evaluation. The performance of the information extraction algorithm was 90% accuracy on average, demonstrating a high level of effectiveness by this approach.

Then, we moved to enhance a new representation for collections, the composition space (see Figure 2). A collection is typically represented as a linear list. This representation makes it difficult to visually compare and connect information from different sources. The composition space is the interactive environment which connects elements visually in a spatial format, using various visual techniques, including compositing [5]. In mixed-initiative composition, in addition to the user’s direct manipulation, system agents automatically generate the composition. This supports participants in directly accessing and browsing information without clicking links. We developed a new control structure for mixed-initiative composition, the *ResultDistributor*, which promotes diversity in the information presented through the composition space, by structuring the prioritization of document processing. The *ResultDistributor* controls the document processing order to be balanced among multiple sources, like a round-robin scheduling algorithm. User evaluation demonstrated that the participants were able to browse more diverse information with the *ResultDistributor*-enhanced composition space. Participants also found it easier and more entertaining to browse information in this modality.

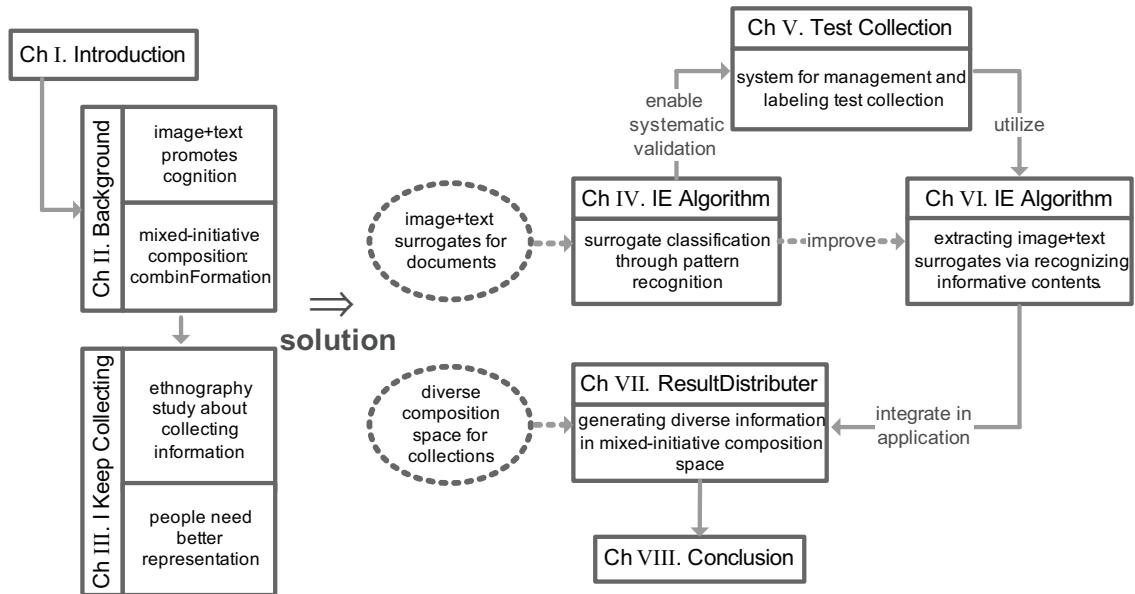


Fig. 3. Concept map: organization chart of the dissertation structure.

The present research connects ideas from cognitive psychology, ethnography, pattern recognition, information retrieval, algorithms, and human computer interaction to develop new methods for visually representing collections. The results are applicable to representing information resources in contexts such as search engines and digital libraries. Representing collections visually, with image+text surrogates and mixed-initiative composition has the potential to enhance people’s everyday experiences of information.

This dissertation is organized as follows (see Figure 3). We begin by presenting the background of this research in Chapter II. Then, Chapter III develops a user study that investigates people’s needs in building and utilizing digital collections. In Chapter IV, we present our surrogate classification algorithm. We describe details of developing significant discriminatory features for the classification, and the experimental results and analysis. Chapter V presents the test collection system that we

developed to collect and label the test data for evaluating our information extraction algorithm, which is presented in Chapter VI. The information extraction algorithm builds upon the surrogate classification algorithm of Chapter IV to automatically identify the informative content blocks within documents, and generate representative image+text surrogates. An evaluation of the algorithm using the collected test data is presented in Chapter VI, as well. Chapter VII shifts from representing individual surrogates to collections, presenting a new control structure for processing documents, *ResultDistributor*. In the final chapter, we draw conclusions on our research.

CHAPTER II

BACKGROUND

In this chapter, we present the background that motivated our research conceptually. We explain the definition of the surrogate, typical representation of surrogates, and some of the related research that develops new representation of surrogates for people. Then, we present the cognition perspective of the image and text representation. Finally, composition space and the combination's mixed-initiative composition for representing collections are presented.

Specific prior work about ethnography, pattern recognition, information retrieval, and human computer interaction research approach is presented in the chapter corresponding to the related research.

A. Surrogate

A surrogate represents an information resource and enables access to that resource [6]. It is a replacement for an original item, which gives some description of the item, and how it can be obtained. One typical surrogate is the Google surrogate, an element of the result set returned by a search query (see Figure 4). People make critical decisions based on these surrogates, such as choosing which documents to browse, and which to ignore. They play a fundamental role in people's process of comparing and choosing documents. During searching and browsing, the surrogate stands between the user and the document; thus, we can say that the surrogate eclipses the document. There are 6.4 billion search queries issued per month in the U.S. alone [7, 8]. 91% of Internet users are located outside of the U.S. [9]. This suggests over 2 billion queries per day, globally. Thus, the representation of surrogates is very significant.

DIVERGENT THINKING

The goal of **divergent thinking** is to generate many different ideas about a topic in a short period of time. It involves breaking a topic down into its ...

faculty.washington.edu/ezent/imdt.htm - 5k - [Cached](#) - [Similar pages](#) - [Note this](#)

Divergent thinking - Wikipedia, the free encyclopedia

Divergent thinking is a thought process or method, which is usually applied with the goal to generate ideas. It is often used for creative and problem ...

en.wikipedia.org/wiki/Divergent_thinking - 18k - [Cached](#) - [Similar pages](#) - [Note this](#)

Convergent and Divergent Learning

The other he termed "**divergent**" thinking. Here the student's skill is in broadly creative elaboration of ideas prompted by a stimulus, and is more suited to ...

www.learningandteaching.info/learning/converge.htm - 15k - [Cached](#) - [Similar pages](#) - [Note this](#)

Fig. 4. Linear list of Google surrogate: A typical surrogate representation.

B. Surrogate Representations

The most common presentation technique for displaying web search results is to show the title, URL, and a short summary of each result. Mostly the summaries show sentence fragments that match one or more query terms. The use of key sentences extracted from the text on destination pages has also been tried with encouraging results for improving web search [10]. Dumais *et al.* [11] explored the use of hover text to present additional details about search results based on user interaction. Paek *et al.* provide more descriptive text by combining a fish eye lens with progressive exposure of page content [12].

More hybrid approaches have been applied for the surrogate representation. Marchionini *et al.* investigated the use of multimodal surrogates for video browsing [13, 14] by comparing users' performance and experience using different surrogate formats for digital videos. Combined surrogates lead to better comprehension and reduced human processing time. Woodruff *et al.* investigated the efficacy of "enhanced thumbnails" as navigational surrogates for documents [15]. They start with a reduced screen shot

of an entire web page. Each thumbnail is annotated with a larger textual “call out,” which indicates the presence of a key phrase from a search result set. Users performed significantly better on search tasks with enhanced thumbnails, than they did with text summaries or plain thumbnails. Our approach, the image+text surrogate, builds on these results. The image+text surrogate is the integration of informative image into a typical text-only representation to make better use of human cognitive resources. The image+text surrogates are formed by transforming document into a small set of concise representations that are easy to understand and think about. Many researches have demonstrated the efficacy of this representation for browsing and navigation. The research reported in this dissertation develops an information extraction algorithm for the automatic generation of the image+text surrogates.

C. Cognition of Images and Text

In the human working memory system, the visuospatial buffer (which stores mental images) and the rehearsal loop used for words are complementary subsystems [16]. They support each other in combined image-text knowledge representations. Glenberg *et al.* have established that the combination of an image and descriptive text promotes the formation of mental models [17], and extends working memory capacity [18]. Carney [19] and Moreno [20] have found that dual coding strategies enhance cognition during educational experiences of digital media. Text disambiguates images while engaging complementary cognitive subsystems. Thus, combining images and text while forming surrogates makes better use of cognitive resources than text alone, and also image and text format was demonstrated to improve navigation [1].

D. Composition Space

From the results returned by search engines, the lists of documents in file directory explorers in computers, and menu of bookmarks provided by web browsers, we can see that the format typically used to represent collections is the list of textual surrogates. Composition is an alternative to lists; literally, it means, “the act of putting together or combining . . . as parts or elements of a whole” [21]. The composition space is an environment where people can compare, connect and relate information visually in a collage form using direct manipulation [5, 22]. With design tools that enable layering, such as changing the elements’ size and color, they can discover new meanings amidst the information. Figure 5 is an example of a composition that represents areas of the undergraduate psychology curriculum using combinFormation.

E. combinFormation: Mixed-Initiative Composition

By adding generative agents to the composition space, the space becomes more dynamic and stimulating. Combining human direct manipulation with generative agents results in *mixed-initiatives*. The present research builds on the mixed-initiative system, combinFormation for developing personal collections of information resources through composition of image and text surrogates [23, 5, 24]. combinFormation is a system that supports people in searching, browsing, collecting, and composing information visually and interactively. The system has been developed in Interface Ecology Lab at Texas A&M University for more than 6 years under the direction of Dr. Andruid Kerne. The development and evaluation of the system has been supported by National Science Foundation.

combinFormation assembles results from multiple search queries. Users are able to collect and compare found information through visual clippings, forming concep-

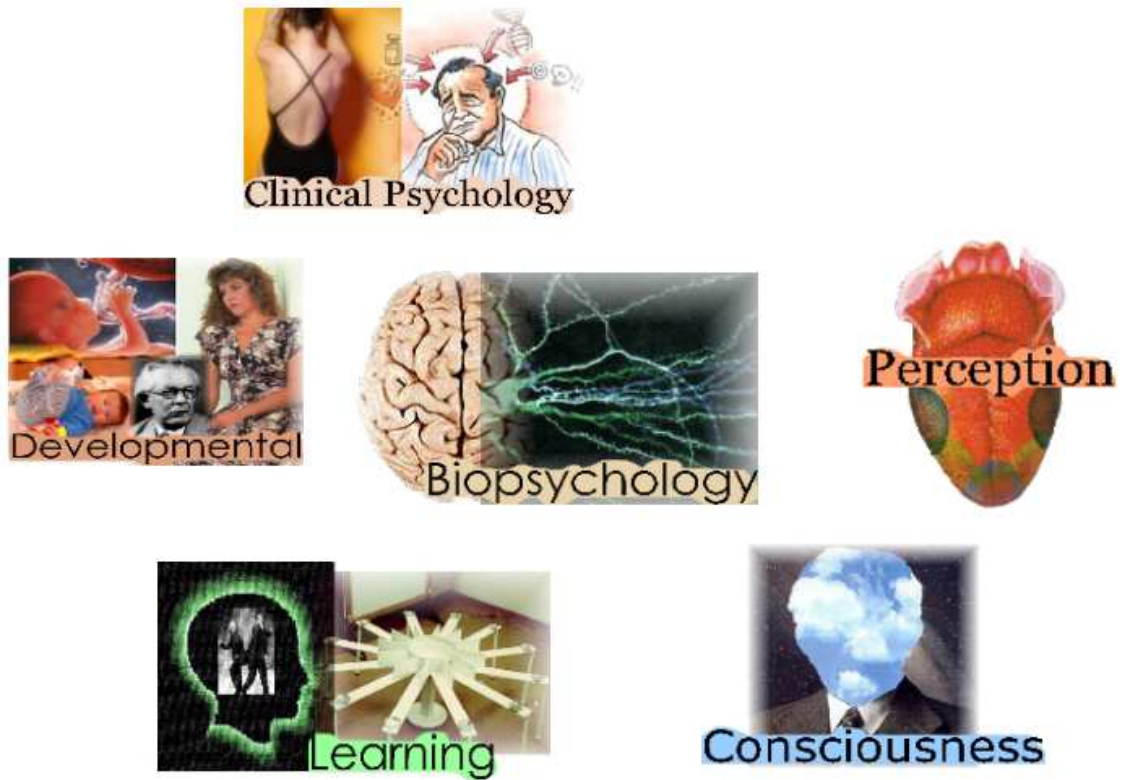


Fig. 5. A composition of image and text surrogates representing sets of information resources for an overview of part of the undergraduate psychology curriculum. Each surrogate is formed by clipping information elements from source containers.

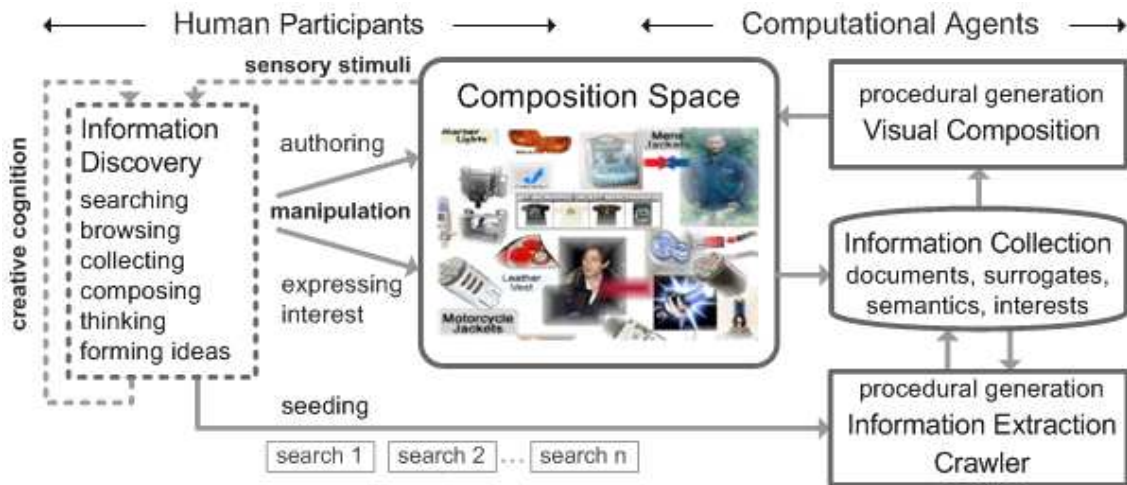


Fig. 6. Mixed-initiative composition: Interactions of human beings and procedural generation of digital information in the mixed-initiatives visual composition space.

tual relationships. Like bookmarks, the clippings function as surrogates for source documents. A fluid interface enables users to express interest (i.e., provide relevance feedback) in each surrogate in-context, with minimal effort. The system agents respond to interest expressions by crawling, retrieving, and presenting relevant information. While interacting with information, users can learn and develop new ideas.

One way to start combinFormation is by mixing multiple queries. Each time a query is entered, another query input box is dynamically displayed [23]. For each search query, users can select a search engine or social information service, including Google, Yahoo, Flickr, or Delicious. combinFormation processes each search by sending the query to the selected engine, obtaining the result set, downloading the result pages, and extracting image and text information clippings. The image and text clippings function as semiotic and navigational surrogates that represent the result documents. The system agent selects these surrogates one at a time, and, over time, combines them visually into a composition space for the user. The user can

concurrently interact with the image and text surrogates in the composition space by rearranging, resizing and changing design. Agent and user actions are interleaved, in a mixed-initiative composition (see Figure 6). Transparent borders, which create visual connection, can be turned on or off for each surrogate. The color of text shading and the font can be manipulated. When the user brushes a surrogate with mouse-over, s/he sees in-context metadata details on demand. S/he can navigate to the source web pages using the navigate tool. While browsing the web, s/he can also drag and drop interesting information into the composition space, and make notes (annotate) using the text edit tool.

This research enhances the existing approach in `combinFormation` by developing an information extraction algorithm that enables generating more meaningful and representative image+text surrogates. Also, the `ResultDistributor` enables the generation of diverse information evenly from multiple sources in the mixed-initiative composition. We evaluated the `ResultDistributor` by integrating it into `combinFormation`'s mixed-initiative composition, and found that the participants were able to browse more diverse information easily in this modality. Before we explain the details of our research approaches, we start from presenting our ethnographic user study that discovered people's experience and needs with collections in next chapter.

CHAPTER III

I KEEP COLLECTING

This chapter presents an ethnographic study to develop understanding of current practices and resulting breakdowns in people's experiences of building and utilizing collections.

Dick is a graduate student in industrial engineering. As he is a research assistant, his work involves writing research papers. He regularly searches for and collects relevant prior work from the Internet and digital libraries. He collects articles and URLs on his own computer. He utilizes this collection regularly. Jane is a visualization lab student. She collects many images and pictures for class work such as animation, and also for fun. Some of these are photographs she has taken; some come from the Internet. She is also a student worker in the university newspaper. She collects images to support this activity, as well. These examples illustrate the contexts in which students are making collections, and provide a sense of the scope of collections and collecting activities addressed by this paper. We define *collecting* as people's practices of putting together archives of information elements, such as hyperlinks, documents, images, audio, and video, with the intention of creating and supporting meaningful, engaging, and useful experiences.

Due to popularity of digital media devices and the abundance of information on the web, a broad cross-section of society becomes more and more exposed to large numbers of digital documents and media elements. People are confronted with the problem of how to keep track of significant elements within the stream of this experience. They begin collecting, and again due to the preponderance of meaningful digital information and media, the collections become larger and larger. This trend is further promulgated by the increasing availability and capacity of inexpensive digital

storage devices.

However, a wealth of information creates a poverty of attention [25]. The disparity between the growing amount of information and media that people are collecting in practice, and their fixed amount of attention, is leading to breakdowns in their collecting experiences. According to Winograd and Flores, breakdowns occur when there is a discrepancy between our expectations and actions, and the world [26]. Breakdowns can serve as an opportunity for learning, because they identify important parts of tasks and activities, and can provoke the articulation of new user needs and design requirements. This section presents research that investigates breakdowns in collecting practices.

The study has been conducted with college students. College students tend to be fast movers in the face of ongoing technological transformation. 81% of them go online. Many of them can scarcely imagine what the world was like way back when people weren't always connected to the net, "Always on" [8]. The Pew Internet and American Life Project, reporting on 2054 students from 27 college and university, says that nearly 73% of college students use the Internet more than the library, while only 9% said they use the brick and mortar libraries more than the Internet for information searching [27]. College students typify the category of *power creators*, which Pew has identified as an important constituency of Internet users [27]. Power creators are twice as likely to engage in content creating activities as other Internet users [27].

This research is to develop understanding of current practices and resulting breakdowns in building and utilizing digital collections. We have investigated the practices of college students by interviewing them, and observing the collections that they build. We also gathered quantitative data about collection building and utilization practices. From this understanding, we will infer implications for the design of new tools to support these processes. The research in this chapter is published in

European Conference on Digital Libraries 2006 [28].

A. Background

Prior studies have investigated the usage of tools for building and utilizing collections in specific media, such as email [29], bookmarks [30, 31], and files [32]. Some studies have offered classifications of user behavior with various collection tools. Malone identified two fundamental strategies in office management: filing and piling [33], focusing on the organization activities. Whittaker and Sidner [34] observed three email management strategies: frequent filer, spring cleaner, and no filer. Balter [29] extended this classification by dividing the no-filer class into folder-less cleaner and folder-less spring-cleaner, depending on whether items are deleted from the inbox on a daily basis. Abrams *et al.* [30] described four bookmark management strategies: no-filer, creation-time filer, end-of-session filer, and sporadic filer. Barreau and Nardi [35] looked at the types of information manage by users, identifying three types based on lifetime and use: ephemeral, working, and archived. They noted the relative importance of ephemeral/working items retrieved by location-based browsing over archived items and the use of search. However, as the information age matures, it seems that the importance of archiving grows.

While each of the previously mentioned works addresses utilization of a single collection medium, Jones *et al.* conducted a study that traverses collecting practices involving e-mail, images, document addresses (URLs), and documents [31]. They investigated various methods people use in their workplace to organize information for re-use. They found that people differ in their collection building practices according to their job position and their relationship to the information. Their study is similar to the present research in its addressing of multiple collection media, as well as in

the number of experimental subjects, and the social proximity of the subjects to the researchers. Boardman *et al.* [32] also collected cross-tool data relating to file, email and web bookmark usage. They found that individuals employ a rich variety of strategies both within and across collection tools, and discuss synergies and differences between tools, to guide the design of tool integration. The data underlined the challenge of the collection tool design by addressing that future design work must take account of the variation in strategies by providing the flexibility to manage different types of information in distinct way. They observed that people usually browse rather than search to find relevant elements in their collections. In addition, they found that the slow-changing nature of hierarchical representations may benefit users by promoting familiarity with the personal information environment. Such familiarity, in turn, supports location-based finding, for which users expressed a clearer preference.

The present research focuses on human experiences of collecting and the role of collections across a broad range of meaning-making activities and digital media. Some prior work has addressed particular media, such as web pages or email. Some has focused on well-defined scenarios regarding information filing, finding, and management. This study investigates processes of collection building and utilization across media and tools through open questions about participants' situated practices, in order to discover how they engage in collecting throughout their everyday activities. We use a hybrid data collection approach, in which qualitative data from open questions is augmented by quantitative data about collection building and utilization.

B. Study Description

To investigate power users' collection building and utilizing practices, we performed a study consisting of interviews of 20 college students. The study brought together

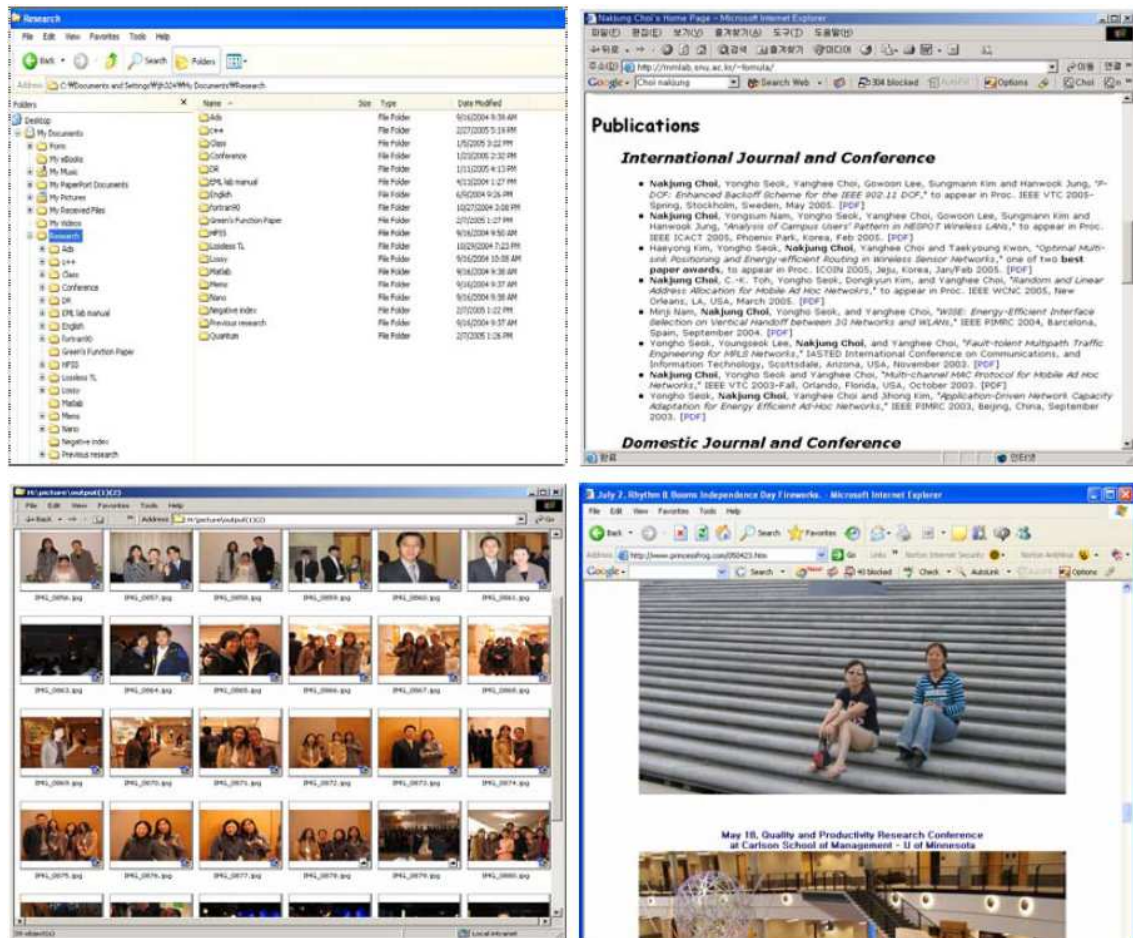


Fig. 7. College students' digital collection examples. Top-Left: a personal paper collection saved in file folders. Top-Right: a published personal collection of papers on the web to share with others. Bottom-Left: a personal image collection in a file folder. Bottom-Right: a published personal collection of images on the web for sharing.

narrative accounts, interview questionnaires, and examples of their digital collections (see Figure 7) in order to investigate how they currently build and utilize collections as part of everyday life. Students were informed that they were participating in a study, and that their responses would be recorded, and anonymously recounted in a research paper.

Participants were distributed by gender and academic concentration. Ten students were male and the other ten were female. There were eight undergraduate students and twelve graduate students. Students' majors were diverse, including computer science, visualization, aerospace engineering, statistics, landscape design, industrial engineering, and history. The interviews were conducted with participants at their offices or homes, so they could show artifacts from their personal computers.

The interviews were semi-structured and open-ended. We did not limit the dialogue to our pre-formulated questions. We also did not place any limits on the media type or representational forms of the collections we investigated. Rather, we considered any type of personal collection. We spent 60-90 minutes with each participant to explore the kinds of collections they made, their processes of using and organizing the collections, the collection tools they used, and their overall experiences of collecting.

While conducting the study, the interviewer was guided by an agenda of relevant research questions:

- To what extent do you think intentionally about your needs for collecting digital information prior to actually doing so?
- What activities are involved in your collection building processes?
- How do you feel about spending time through collection making processes?
- How many elements are in your collections?

- Which tool(s) or mechanism(s) do you use to build collections?
- How often do you make / refer to / organize collections?
- What types of inconveniences and breakdowns do you encounter during building and utilizing digital collections?
- What are your strategies for coping with breakdowns in your experiences of building and utilizing collections?
- What are your suggestions for future collection tools?

We recorded and screen-copied examples of collections participants built, and took notes of interviews. After each interview, participants filled out a survey questionnaire.

C. Study Results

We analyzed the study data in terms of the distribution of activities, significance, type, and quantity of information elements involved, as well as the kinds of mechanisms people used for building and utilizing collections. We also investigated their frequency of involvement in collecting. Quantitative and qualitative data and its analysis will show participants' collection building and utilizing practices and behavior. Quantitative data is collected by participants' self reports of their personal collection practices. We value more the participants' self reporting data, because what they understand about their collection is at least as significant as what they actually have in their digital devices. When participants are utilizing and referring to their collections, they interact with the collections based on their understanding.

1. Collection Building and Utilizing

We looked at collection building and utilizing practices in terms of the stance participants brought into the process of collecting, the patterns and expectations that occurred in these processes and the ways in which users perceived success and failure.

2. Intention and Need

Participants were asked whether they thought about the need for collecting prior to engaging in processes of seeking digital information. All participants expressed awareness of a personal ongoing deliberate intention and need to be involved in collection building and utilizing practices.

3. Activities and Significance

The participants reported collecting digital media materials that support a range of personal and work-related activities. The personal media included photographs taken by themselves and friends, as well as popular media elements such as music, movie star pictures, and art images. As the subjects were students, their work is learning and research, so the materials here included class notes and research papers. Students whose majors are related to design collect many image files as part of their school work. From this data, we see that the participants' collecting activities are conducted in relationship to the span of significant activities in their lives.

4. Frequency and Time Period

One hundred percent of participants report that they build and utilize collections regularly. Of these, more than half utilize collections more than one hour per week. In more detail, 18% of participants said that they spend more than one hour per

day on collection building; 10% spend one hour per day engaged in the collection process; 27% said that they spend more than one hour per week and less than one hour a day; while 27% spend one hour a week; and 18% of participants spend one hour per month. However, participants do not have a specific time frame scheduled for collection building and utilizing. It is something they do spontaneously, as part of a range of tasks and activities (P3: *“I build and utilize collections regularly, and I engage in this process during spare time and while I am taking rest.”*).

5. Worthwhile or Useless

Participants were asked how they feel about spending time on collection building and utilization. 46% of participants said that they experience the process as meaningful and worthwhile. 18% of participants answered that they find it somewhat meaningful. 9% of participants answered that their experience is neither worthwhile nor useless. 27% of participants said that they experience collecting as rather useless. Those participants who answered rather useless said that they nonetheless continue to engage in the collection building process; they experience it as necessary and meaningful initially, but after a while, their engagement seems to be performed in vain. They said that a collection is not worthwhile if they do not utilize it well, and they seldom utilize most parts of their collections because of the huge volume of collected information.

6. Collection Types

All participants said that they build image, music, and/or movie collections. The sources of the images are from digital cameras, camera phones, and the Internet. Twenty-two percent of participants have 50-100 images in their collection; another 22% keep 100-500 images; while 56% keep more than 5000 image collections. Participants said they mostly obtain music from music downloading services or their

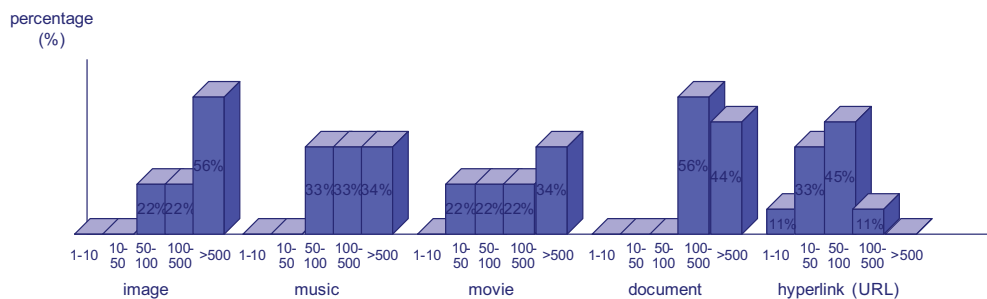


Fig. 8. Number of elements participants collect, by media type.

friends' collections. Thirty-three percent of participants keep 50-100 music files, 33% keep 100-500 files, and 34% keep more than 500 music files in their collections. Movie files are obtained through similar means, such as downloading services or creation with a video camera. Twenty-two percent of participants keep 10-50 movie files; another 22% keep 50-100 movie files; another 22% keep 100-500 movie files; while 34% keep more than 500 movie files in their collections (see Figure 8).

Participants also collect documents such as Word files and PDF files. 56% of participants keep 100-500 documents; 44% keep more than 500 documents in their collections. They also collect web documents in the form of hyperlinks (URLs). 11% of participants keep 1-10 URLs, 33% keep 10-50 URLs, 45% keep 50-100, and 11% keep 100-500 URLs in their collections. Compared to the other media collections, participants keep fewer URLs, because web documents are easier to search for.

7. Collection Mechanism

In terms of what is stored, there are three ways to build digital collections: (1) save the files themselves; (2) extract some parts from files and save only those parts; (3) save the location of files. Participants use whatever tools and structures are at hand to build their collections; for example, files, folders, bookmarks, and e-mail.

All participants said that they make file folders for file collections. There are also

within-file collections, in which small elements of information from diverse sources are gathered into a single file. Participants said that they used Excel, Word, Photoshop, and Notepad to build this type of within-file collection. They used drawn lines, tables and newline characters (vertical whitespace) to spatially distinguish elements in a within-file collection. When participants save URLs of web pages, they usually use bookmarks, but they also use e-mail, so that those URLs can be utilized from the other computers (P9: *“I am not using bookmarks at all. Instead I keep URLs in my email because I use three computers; my office computer, my home computer, and my laptop, so I can look at important URLs from any of my computers.”*).

8. Levels of Engagement with Collections

We observe that in general, people collect information and media with the intention of later referring to the collected elements for use. Sometimes, they actually get to this process of referring. Further, sometimes, with collections that are important, they take steps to organize the form of the collection. Referring and organizing are aspects of collection utilization.

While participants accessed the Internet daily, their activities of selecting elements to add to their collections, referring to the collections, and organizing them occurred less frequently (See Figure 9 Top). The frequency of these activities can be categorized in three tiers. All of the subjects accessed the Internet daily. At the same time, 43% of them engaged in collection building and referring on a daily basis, while 36% did so on a weekly basis, and the remaining 21% engaged in such activities monthly. The difference between Internet access frequency and collection building/referring frequency was statistically significant [$F(2,26) = 3.67, p < 0.01$]. While distribution of the participants' collection building frequency and collection referring frequency were the same, these distributions are independent and do not

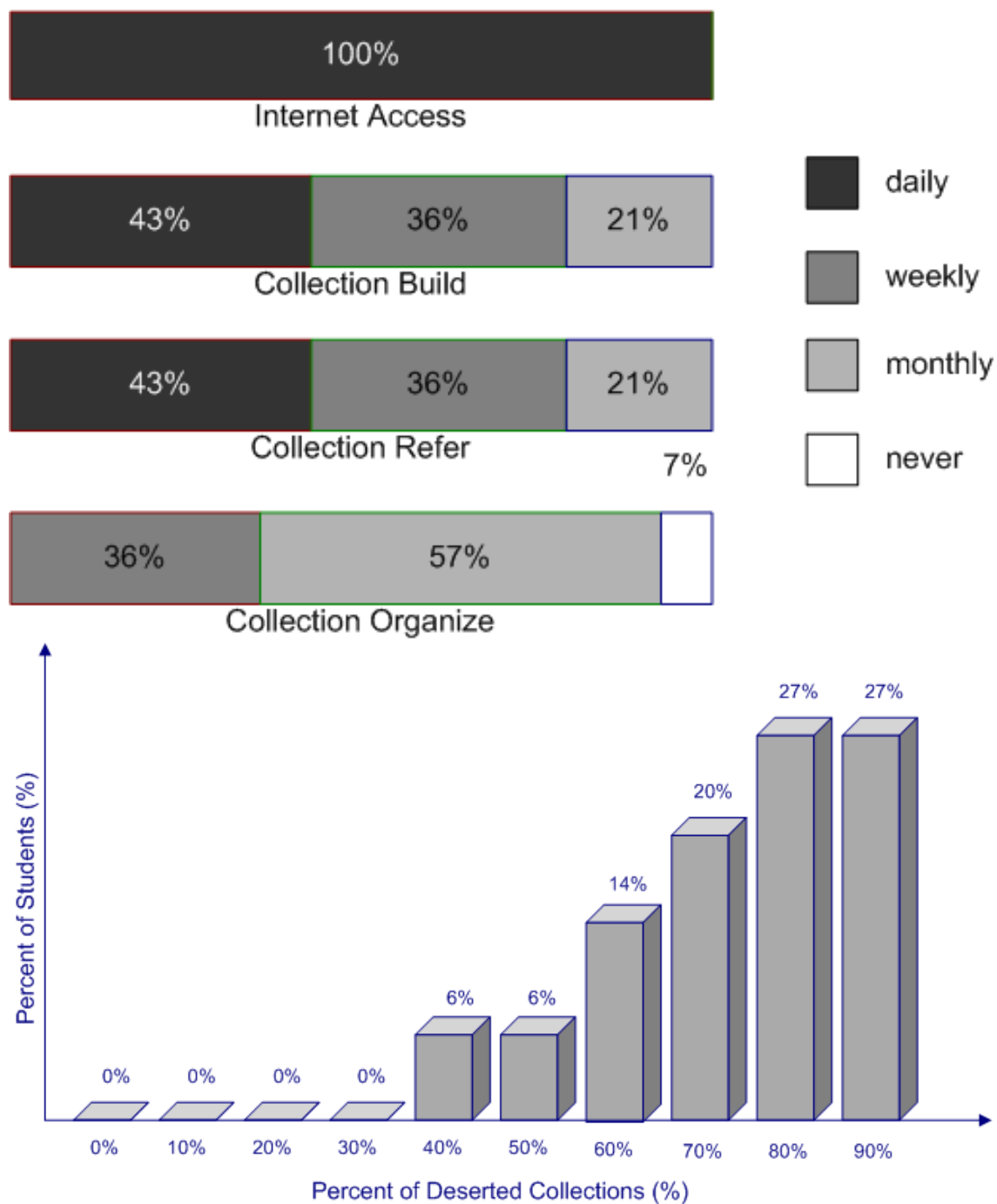


Fig. 9. Top - participants' Internet access and collection building, referring, and organizing frequency; Bottom - rate at which participants' collections are unutilized and abandoned.

necessarily refer to the same participant. The third tier of engagement with collections is to organize them; 36% of the subjects did this weekly, 57% did it monthly, and the last 7% reported they never did it at all. The last group corresponds, for example, to Abrams, “no-filers” [30]. The frequency of engaging in collection building/referring was again greater than that of collection organizing in a statistically significant manner [$F(2,26) = 3.45, p < 0.002$]. This shows that people refer to their collections as much as they build the collections, but they rarely organize their collections.

9. Collection Sharing

The study data shows that participants share their collections with other people, and also across several computers. 85% of participants said that they have their own blogs or personal web sites and publish some of their collections to share with others. These published collections may in turn function as source materials for others’ collection building processes.

As mentioned above, one participant (P4) keeps URLs in email in order to access them from different computers. All participants said that they use several computers in different places. Participants use portable devices to carry their digital media materials or store them in network accessible spaces in order to share among different computers and as well as with others.

10. Breakdowns in Collection Practice

We investigated discrepancies between participants’ expectations, and their experiences in practices of collecting. Our goal in identifying these breakdowns is to articulate user needs and design requirements. The most common breakdowns that participants experienced during the present study arose during their practices of referring, organizing, and finding things in their collections (P15: “*I initially made*

URL collections using bookmarks without any folder structure and renaming. Later, I had trouble finding a specific URL in it, so I deleted all my bookmarks and made folders with renaming. After this experience, I became more cautious about adding and renaming URLs to the collection.”). They said that they initially didn’t have trouble finding elements in collections they built, but as time elapsed after collection building, it became more difficult to remember what is in the collections, and where. Recall, a problem of limited human attention, becomes a problem (P12: *“I had really important data in my collections, but I cannot find it! Could you make a program for me?”*). As the set of collections they own grows larger, it becomes difficult to remember all of them. Even though they sometimes don’t have any clue of where the elements are, they said that they start browsing their collections first rather than searching. When they don’t find the elements in the expected location, they use a search tool (P13: *“I seldom organize my collection very well, so I went through all folders one by one sequentially trying to find a certain file. Sometimes, I forgot what I saved, so I searched the web instead of the collections, and saved the same thing again.”*). However, they may not even remember what to search for.

As mentioned above, 27% of participants said that collection building is somewhat useless because most parts of their collections are not utilized, and thus abandoned. Participants were asked what percent of their digital collections remain unutilized. At least 40% of the participants’ collections are abandoned (See Figure 9 Bottom); 27% of participants said that 90% of their collections are abandoned; another 27% of participants indicate that 80% of their collections are abandoned; for 20% of participants 70% of collections are abandoned; 14% of participants have a 60% abandonment rate; 6% of participants have 50% abandoned collections; another 6% have 40% abandoned collections. Nonetheless, participants continue to engage in collecting (P4: *“Even though I am not using most of my collections and I sometimes think what*

I've built is useless, I keep building collections.”).

The participants initially build their collections with the intention of using them later. However, most collected material is not utilized because of trouble remembering and finding what has been collected. They lack effective means for referring to their collections. Collections are abandoned not because the information and media they contain are useless, but because of breakdowns in utilization practice.

11. Reasons for Collection Building

Participants were asked why they still build collections even though they do not utilize most parts of them. Like P14 (“*Wow, I realize that I am not using most parts of my collections, around 90%*”), they are often unaware that they are not utilizing most of what they collect. However, all participants still build collections from some sense that they will need the collected information elements later (P6: “*I want to save time on searching when I need a document in the future. That is my main reason for continuing to build collections.*”). They collect media files to enjoy and also to share with others. Participants collect information that seems meaningful, useful and needed. They collect media that seems fun, unique, and consonant with their personal tastes. They make collections not for the definite promise of later utility, but from some intuitive sense of meaning and value.

12. Using Semantics to Represent Collections

Through the study, we observed that participants create semantic structures to organize their collections using any available affordances. They build their own structures for meaningfully representing their collections for usage later.

13. Developing Informal Metadata Schemas

All participants said that they make hierarchical directory structures to organize and manage their collections. They make folders based on contents, dates, semantic identifiers related to tasks or activities, or other categories that are somehow significant to them. Participants said that folder structures are created and changed because collections are added and deleted continuously.

Participants said that they rename files and file folders using metadata such as date, location, title, or author in order to help find them later. Renaming is important for search also. They seek to remember which words they used to rename files, in order to reuse them later when they browse and search their collections. Several participants mentioned strategies other than renaming for keep tracking of collected material. For example, they create index files inside of folders so that they can know what they contain (P6: *“Inside file folders, I make a ‘readme’ file to look at it later. This will help me to remember what the collection is about. In the individual file, I rename the file, and in addition to that, I put an explanation about the content in the first line.”*).

We identify participants’ practices such as renaming elements and creating hierarchical folder structures for representing important and large collections as the development of informal metadata schemas. They found ways to develop informal metadata schemas even in the absence of tools that support extensible field creation. They used the single accessible field afforded by existing tools that is the file or link name, to store the metadata. This practice was mostly spontaneous, occurring without an ontological plan. It was conducted informally and incrementally, as a series of situated actions [36]. This is an example of incremental formalism [37].

14. Suggestions

Participants were asked what new functionalities would be helpful in tools for collection building and utilization. Categories were not specified. Participants could mention whatever was on their minds. Participants' suggestions addressed areas such as collection utilization statistics display, filing assistance, and collection privacy support. They wanted help in renaming their collection materials in order to make the structure consistent, to make it easier to find materials later. They also asked for cues such as a 'visited count,' which shows how many times the owner read the file, in their collection representations and search and browsing environments to support finding specific materials. They liked the way desktop search is moving to assist collection utilization, however, they wanted their private files to be processed differently (P13: *"I have a big paper collection, but it is hard to find the paper I need when I need it using search tools supported in Windows. I tried Google desktop search, and it is pretty good, but one time I was a little embarrassed because a file that I wanted to keep private was retrieved as a search results when I was with my friend"*).

D. Implications for Design

Study participants invest substantial personal effort and resources into processes of building and utilizing collections. Their persistence in collecting in spite of breakdowns conveys the sense that they need to keep collecting to support a range of activities that span personal and work-related parts of their lives. In this section we examine participants' engagement with collections and the needs they express, and extrapolate from these, while considering human cognitive facilities and emerging technological capabilities. The result is to derive implications and ideas for designers of systems that support collection building and utilization.

The data shows that participants' breakdowns were centered in processes of collection utilization. They had trouble finding specific elements in their collections, and even though they built collections of elements that were useful, most of them are not utilized in the relevant context because of limited human attention and memory. They forget what to look for and where. Abandoned collections consume disk space, and more importantly, human attention during browsing, which is people's first choice for how to refer to collections.

We propose prescriptions to address breakdowns discovered in this study. Since the discovered breakdowns generally involve limitations of human understanding of collections, the prescriptions involve making better use of individual visual cognitive resources, sharing collections, and the definition of collection semantics. The first prescription addresses breakdowns that involve forgetting what has been collected, by using representations for collections that better cue human memory. The next proposed solution is based on ambient displays that use peripheral attention and changes over time for individual and collaborative interaction with collection visualizations. Other user needs that result from analysis of the breakdowns involve distributed tools for collection sharing, and the automatic generation of metadata schemas.

We can take steps to help people keep track of their collected information, by making better utilization of human memory capabilities. It is a well-accepted principle of cognitive science that in the working memory system, the visuospatial buffer, which stores mental images, and the rehearsal loop used for text are complementary subsystems [16]. Thus, dual coding strategies that represent the elements stored in a collection with images as well as text will improve memory utilization [16, 19], and contribute to helping people find elements while browsing. Thus, we can provide users with tools that support them in developing and generating visual index representations of their collections, which integrate images and text. These representations will

be easier to remember, promote recognition, and facilitate the formation of mental models [17]. Since collection representations function as visual communication, either from a user to her/himself or between users, visual design principles must be applied.

Developing representations during collection-building and explicit organization activities is one solution. But people don't have sufficient attention to work on representing all the elements of their collections. Another prescription develops peripheral ambient visualizations that gradually display elements from collections over time. Ambient visualizations use time as a dimension in collection visualization. They can represent personal and group collections, engaging human attention without requiring it. Ambient or mixed-initiative visualizations can be deployed on a dedicated display, as a screensaver, or in a window that receives human attention only periodically. The set of collections that get visualized can be specified explicitly by users, and/or by an agent like mixed-initiative composition that uses clues, such as people's interest. For example, a large display in a collaborative environment such as a research lab or departmental work area can visualize collected materials that represent information relevant to current projects and research. This method can jog memories and promote serendipity, to facilitate individual and collaborative utilization of meaningful, useful and important elements in collections. Affordances that enable privacy will be required.

Additionally, we have seen that sharing with others is an important motivation for peoples' collecting practices. People utilize and collect information on multiple computers and devices in different locations. This can cause access problems, when the person is in one place, and the needed information is somewhere else. One initiative that addresses this is 'del.icio.us', which supports URL collection sharing. del.icio.us enables users to tag URLs while collecting. It shows the metadata that others have used, and enables social browsing through these relationships. We believe this is a

start for sharing collections and their semantics. New collection tools need to consider people's social and distributed collection-sharing intentions and enable collecting actual objects as well as references, while considering accessibility and privacy. Deeper semantic structures than single tags will also add value. These functionalities need to be integrated with editing, saving, browsing, and searching in order to best use limited human attention.

The findings of Czerwinski *et al.* [38] and Marshall *et al.* [39] are similar to what we have found in this ethnographic study. They mentioned that recording, creating, receiving, storing, and accumulating digital materials is easy, but managing and using them sensibly is difficult, especially as time passes and their immediacy fades. It is obvious that many people are confronted with the problem of referring to specific information in their personal collections, especially as the collections become larger and older. The implications for design that we derive are quite different from what they present. They focused more on software that controls distributed storage. They develop heuristic notions of the value of information. They also mentioned visualization and representation, but they didn't address what kinds of visualization will help to address the breakdowns. This research develops more concrete idea to address breakdowns that humans experience while utilizing personal digital collections, by providing better representations for individual information elements and collections.

Our study participants display tenacity in their involvement in processes of collecting. They explicitly express the intention and need to be involved in ongoing practices of collecting. They collect digital media materials involved in a broad range of activities, spanning personal and work relationships, which make up their everyday experiences. Their collection artifacts directly signify, relate to, and support these activities. Thus, collections and the process of collecting, itself, play important roles in how people create meaning in their lives.

Participants engage in collection building and utilizing activities regularly, even though it is not mandatory, and even though problems arise in the user experience. They keep collecting in spite of breakdowns. Better representations can help support these processes, by making better use of human attention.

The following chapters present our research approaches to better representations of digital information. The first approach follows the prescription of presenting documents with the integrated image and text surrogates by developing an algorithm to automatically generate such representation. The second approach follows the prescription of ambient visualization by enhancing the generation of diverse information in the mixed-initiative composition space with the ResultDistributor. The next chapter starts by explaining our first algorithmic approach to forming the image+text surrogate generation.

CHAPTER IV

INFORMATION EXTRACTION ALGORITHM: SURROGATE
CLASSIFICATION THROUGH PATTERN RECOGNITION

People require better representations of documents. They have limited time and attention to find documents in their collections. To help people negotiate the cognitive load inherent in managing large collections of documents, we develop an algorithm for automatically transforming documents to form these representations. The image+text surrogates may be formed by breaking a document down into a set of smaller elements, each of which functions as a *surrogate candidate*. While processing these surrogate candidates from an HTML document, relevant information may appear together with less useful *junk* material, such as navigation bars and advertisements.

This section develops a pattern recognition based approach for eliminating junk while building the set of the image and text surrogate candidates. The approach defines features on candidate elements, and uses classification algorithms to make selection decisions based on these features. For the purpose of defining features in surrogate candidates, we introduce the Document Surrogate Model (DSM), a streamlined Document Object Model (DOM)-like representation of semantic structure. Using a quadratic classifier, we were able to eliminate junk surrogate candidates with an average classification rate of 80%. By using this technique, semi-autonomous agents can be developed to more effectively generate surrogate collections for users. The research in this chapter is published in ACM Document Engineering 2007 [40].

A. Background

Prior research has developed valuable methods for modeling web page documents, defining useful feature sets, and using the features to recognize structures within

these documents. EXALG [41] is an algorithm that extracts structured data from a collection of web pages with a common template. EXALG first discovers the unknown template that generated the pages and uses the discovered template to extract data from the input pages. Arasu *et al.* developed two novel concepts, equivalence classes and differentiating roles, to discover this template [41]. Pages are grouped into sets of equivalent pages based on the presence of common patterns in HTML structure. EXALG constructs a template based on the equivalence classes of multiple pages from each site. EXALG works well for many sites and pages, but there are several limitations. One is that it requires a large amount of space to save the templates. Additionally, EXALG cannot model web pages for which a sufficient number of equivalent pages do not exist.

IEPAD [42] is a system that automatically discovers extraction rules from web pages. IEPAD can automatically identify a record boundary by repeated pattern mining and multiple sequence alignment. The discovery of repeated patterns is realized through data structures called “Practical Algorithm to Retrieve Information Encoded in Alphanumerics” (Patricia, or PAT) trees. A PAT tree discovers patterns in the encoded token string, so it only can see patterns of some parts of a web page. Therefore, this technique is applicable to the extraction of data from highly regular documents with repeating structures, such as search result pages, as evidenced by the experimental collections used in [42]. In the present work, we also define a tree structure (called a Document Surrogate Model), to find tag patterns, but our method is different because it can model patterns in the overall structure of the page.

InfoDiscoverer [43] partitions a page into several content blocks according to the HTML tag `<table>` in a web page. Based on statistics on the occurrence of table tag features in the set of pages, it calculates an entropy value for each feature. The entropy of a content block is defined according to the value of each feature within that

content block. Lin *et al.* found that each page consists of some informative content blocks that can function as distinguishing parts, whereas other non-distinguishing content blocks are more or less the same throughout certain page subsets. We agree that document content blocks can be usefully identified by HTML tag patterns, but it strikes us that the method of identifying content blocks by defining features only through `<table>` tags can be improved upon. We construct content blocks based on a larger set of HTML tags, so that we can identify patterns in a larger class of pages. This is especially important due to recent changes in the way that web pages are authored. Currently, developers often use *Cascading Style Sheets* (CSS) [44] as well as tables in order to structure formatting.

Rowe *et al.* [45] investigated the automatic identification of advertisements within web pages. They performed several experiments to validate various techniques that identify advertisements. They developed a set of features for both image ads and associated texts. In our work, we borrow some of these concepts, such as the presence of a difference between the Internet domain of an image and that of its containing web page.

The present research addresses a problem similar to but different from this relevant prior work. Unlike EXALG [41] and IEPAD [19], we need to be able to process all sorts of HTML documents, not just highly structured ones, such as templated web sites and search engine result sets. We build on [45], extending the definition of junk beyond advertisements. Further, we develop the Document Surrogate Model specifically to represent the structural relationships among surrogate candidates within a document.

Further, a number of the approaches reviewed above attempt to extract information automatically without any human input [41, 43]. As we have seen, purely automatic information extraction has limitations. The applicable scope tends to be

limited to certain web sites. It is difficult to extract information once the style of HTML documents change. In order to extract information from large and diverse collections of documents, it is necessary to utilize human cognitive feedback in collecting training data that can be used later by procedural classifiers to build models of junk candidate surrogates. This is the essence of our pattern recognition techniques, a semi-autonomous process. In our preliminary technique, we define the features for identifying surrogates, the procedure for gathering training data, the algorithms for classification, and the results produced.

B. Surrogate Features

In a general pattern recognition approach, feature sets are constructed to measure certain properties of the data. Sample data can then be represented in a Euclidean space by using the various values of the specified features as coordinates. However, there is no known automatic method for deriving a good feature set; the most reliable metric for feature set performance being classification rate. Feature set determination is a critical part of this work. In choosing our features, we have built upon the work of Rowe *et al.* [45], and EXALG [41], but have also designed new features for the purpose of increasing the separability of the data in feature space.

Our feature set is heavily dependent on tag patterns, the nested set of Document Object Model (DOM) element tags, which contextualize the structured markup of text within a document. Tag patterns are useful for locating "junk," because junk elements are often found in similarly structured regions within HTML documents. For example, advertising companies supply their advertisements using consistently structured HTML tags. Navigational toolbars also tend to be formed with repetitive markup. Unfortunately it is impossible to simply write rules to describe the tag

patterns for junk. It is necessary, instead, to form statistical models from real world data in which humans identify junk in context.

Tag patterns are a vector of the tags surrounding a given element within the HTML document. For example suppose we are given the following HTML code.

```
<body> To Do List

<ul>

<li>Do branch merge</li>

<li>Fix bug id #10 in release 1.1</li>

</ul>

</body>
```

A three-element tag pattern for "Do branch merge" would look like something like this.

```
<li><ul><body>
```

If we were to construct a tag pattern for the second item in the list the same tag pattern would result.

The Document Surrogate Model (DSM) serves as a structural mechanism for constructing features based on tag patterns. Each tag in the pattern is a feature, and is represented as a dimension in the feature space. In constructing a feature set for surrogate candidates, we have drawn from the prior work but have also added DSM-based patterns. We have also added new features based on our general experience with authoring, reading, and examining the source code for web pages. We have constructed two feature sets: one for images and one for textual elements, as summarized in Table I.

Table I. Document surrogate features.

Type	Features
Image	width, height, aspect ratio, alt string length, image name length, image hosted in same domain, ascending 6 tag patterns
Text	length of text, number of non alpha-numeric characters, ascending 8 tag patterns

For image surrogate candidates, we consider image width, height and aspect ratio as per Rowe *et al* [45]. We also consider the alt attribute of the img tag. We use the length of the alt string, because in general, advertisers do not make a practice of utilizing them substantially. For example, on the home page of cnn.com, for all advertisements, the alt attribute = “Advertisement”. Lastly we also include a Boolean feature that indicates whether an image is hosted in the same domain as the document or not, since advertisements are usually hosted on outside domains [45].

For text surrogate candidates, it has been our experience that strings containing a large number of non-alphanumeric characters are usually non-informative. We have also noted that advertisements and navigation elements tend to be short in overall length. For this reason, we also consider the total text length as a feature, since longer text elements tend to be more representative of document content.

C. Document Surrogate Model

The Document Object Model (DOM) is a tree-structured representation of a document [2] that represents markup and text. We introduce the Document Surrogate Model (DSM), a streamlined document tree, in which the significant leaves are surrogates, instead of text nodes. This tree contextualizes surrogate candidates in the

document structure in which they were authored. HTML is parsed to form the DSM, which in turn is used to facilitate the extraction of tag pattern features. The DSM is formed to facilitate representation of the important meanings in documents, and manipulation of such representations.

The DSM is not a complete parse tree; it focuses on capturing the structural elements of the document that are significant for the purpose of surrogate identification, classification, and contextualization. Some markup elements, such as text styles, are discarded, enabling some text and markup DOM nodes to be merged. Additionally, surrogate candidate nodes contain references to their parent nodes in the DSM, thus enabling easy and quick discovery of structural relationships between surrogate candidates.

The most significant departure from the DOM is the use of surrogate candidates rather than text nodes. In a DOM, the content or "text" of the document is completely represented by the text nodes. Extracting just the text nodes from a DOM results in the complete text of the document, but without any of the document structure. From the perspective of representing the set of meanings in a document, it makes sense to consider images as part of the "text" of a document. To do this, we extend the notion of a "text node" to function conceptually, beyond the raw markup structure, by including images, or any other media format.

The complete set of surrogate candidates serves as a representative replacement for the complete "text" that is available in the DOM. The DSM's complete set of surrogate candidates may also be filtered, so as to focus its representation of the meaning of a document. Thus, extracting all the surrogate candidates from a DSM does not necessarily result in the entire "text" of the document; it may be a shorter representative "text" that gives a fair impression of the complete "text." Each individual surrogate node is not expected to be representative of the entire document,

although the intention is that some collection of these nodes will be able to serve as a representative document surrogate.

Surrogate candidate nodes can be formed in numerous ways. For example, `combineFormation` begins constructing the DSM at document parse time. The general approach is to break the complete "text" into small chunks. Then, based on a number of heuristics, chunks are discarded, combined, and otherwise manipulated to create surrogate candidates. Text formatting, script code and stylesheet blocks are discarded as non-informative text. Surrogate candidates often span across multiple HTML tags, combining chunks of the document "text." In this way, some DOM nodes are effectively merged. The resulting surrogate candidate uses only a single parent node. Less meaningful chunks of the "text" are discarded, thus making the final set of surrogate candidates more summative and less of a complete representation of the document "text".

Once the DSM has been constructed, the task is then to somehow use the information present in the structure to perform a second pass of filtering, based on surrogate tag pattern features. The DSM facilitates this task by making the extraction of tag patterns a trivial operation. After the DSM is fully created, it is easy to walk the tree from a given surrogate candidate and determine the tags of its parent and children. Unlike with a standard DOM, there is no extra information that would require further processing at this stage.

Since the problem we are dealing with is candidate surrogate selection, it makes sense to use structures in which surrogate candidates are first class objects. All document data that does not pertain to the surrogates, or the important document structure in which they reside, is removed. As a result, there is far less clutter to deal with when extracting feature information.

D. Pattern Recognition Approach

The pattern recognition approach begins by employing Principle Components Analysis to reduce the dimensionality of the training data so that we, as researchers, can see how the features are contributing to that data’s separability. Next, a pattern classifier is used to classify encountered data points based on a model of the training data. Finally, a cross-validation method maximizes our utilization of the data by rotating which data is used for training, and which for validation.

As described in Table I, our feature space is 12-dimensional for image surrogates and 10-dimensional for text surrogates. Due to the dimensionality of the feature space, it is difficult for a human to see the underlying structure in the data. It is necessary to see this structure in order to define a pattern recognition apparatus that is suited to the data. For this reason, we employed Principal Component Analysis (PCA) to project the data onto a two-dimensional subspace [4].

For those that are unfamiliar with this technique, geometrically, PCA can be thought of as a rotation of the axes of the original coordinate system (basis vectors) along the directions of maximum variance in the data. These directions define a new set of orthogonal axes, which are ordered by decreasing amount of variance in the original data. Dimensions in the resulting space that account for less variation can be discarded, resulting in a low-dimensional projection that retains only the highest variance dimensions. 2D PCA scatter plots are an effective tool for visualizing the structure of high-dimensional data, even though the resulting projections cannot be directly interpreted in terms of the measurement units of the original feature space.

1. Pattern Classifier

A quadratic classifier was chosen for the automatic classification of surrogates. The quadratic classifier assumes that each class (i.e., junk vs. non-junk) is normally distributed with mean μ_i and covariance matrix Σ_i ($i = \{\text{junk}, \text{non_junk}\}$):

$$P(x|\omega_i) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu_i)^T \Sigma_i^{-1}(x - \mu_i)\right) \quad (4.1)$$

where x is the feature vector associated with a given surrogate. Following the Maximum A Posteriori principle [46], surrogate x is classified according to the decision rule:

$$x \in \begin{cases} \text{junk} & \text{if } P(\text{junk}|x) > P(\text{non_junk}|x), \\ \text{non_junk} & \text{otherwise.} \end{cases} \quad (4.2)$$

where $P(\text{junk}|x)$ and $P(\text{non_junk}|x)$ is the probability of a surrogate being junk or non_junk, respectively, given the feature vector x . These functions are also known as posterior probabilities because they define the likelihood of an event (e.g., junk surrogate) after measurement x is taken. Applying Bayes rule, the decision rule in (4.2) can be expressed as:

$$x \in \begin{cases} \text{junk} & \text{if } \frac{P(x|\text{junk})}{P(x|\text{non_junk})} > \frac{P(\text{non_junk})}{P(\text{junk})} \\ \text{non_junk} & \text{otherwise} \end{cases} \quad (4.3)$$

where $P(\text{junk})$ is the frequency of junk surrogates, also known as the prior probability, and $P(x|\text{junk})$ is the density of examples for the junk class, which by equation (4.1) we assume to be normally distributed. $P(\text{non_junk})$ and $P(x|\text{non_junk})$ are defined similarly. Merging equations (4.1) and (4.3) and taking natural logarithms:

$$x \in \begin{cases} \text{junk} & \text{if } g_{\text{junk}}(x) > g_{\text{non_junk}}(x) \\ \text{non_junk} & \text{otherwise} \end{cases} \quad (4.4)$$

where:

$$g_i(x) = -\frac{1}{2}(x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i) - \frac{1}{2} \log |\Sigma_i| + \log P(\omega_i). \quad (4.5)$$

To build a quadratic classifier one need only compute the mean μ_i and covariance matrices Σ_i of each class from training data, estimate the prior probabilities from the expected frequency of junk and non_junk surrogates, and plug these parameters into equations (4.4) and (4.5).

2. Cross-Validation Method

The performance of the quadratic classifier is estimated by means of k-fold cross-validation [46, 4]. In this approach, the dataset is divided into k non-overlapping subsets (or folds). For the i-th fold, data from the remaining k-1 subsets is used as training data to estimate the model parameters in equations (4.4) and (4.5), whereas data from the i-th subset is used as a validation set. In this way, each example in the dataset is used once for validation and k-1 times for training, making the best use of all the data available. 5-fold cross-validation, which corresponds to a (80/20) split, was used for all experiments.

E. Experiments

1. Datasets

We constructed three types of datasets to validate our surrogate classification approach. In each of the three cases, the data consisted of two classes: “junk” surro-

gates and “non-junk” surrogates. The first dataset, which we refer to as the *Structured Collection*, consists of sites selected by the experimenter based on their informative value, or by carefully crafted Google search terms. These sites tended to be well structured and maintained, which suggests that they are automatically generated by publishing systems that utilize templates. The Structured Collection consists of three primary types of sites, news sites (e.g., cnn.com), EverQuest II sites, and travel sites for Costa Rica and Venezuela. This gave us a fair sample of sites that were structured to cover specific topics. The second set, referred to as the *Non-Structured Collection*, results from doing broad Google searches on a set of general terms. These sites are likely to be small web sites or personal web pages with varied design patterns. The Non-Structured Collection contains a wide range of sites, combinFormation [23, 5] was seeded using Google searches on the following terms: cars, research, collections, fashion, personal web sites, travel, about me, fun, health, and gaming. Combining the Structured Collection, and the Non-Structured Collection made the third dataset, the *Complete Set*.

The Structured and Non-Structured Collections were constructed with the help of a human experimenter working with a modified version of combinFormation. Using the cut tool in combinFormation, the experimenter could then manually label the visual surrogates as “junk” or “non-junk”. The resulting feature vector, class name and the surrogate’s URL were then saved to disk. Overall the complete set included 1,496 different samples from 631 different pages over 142 different domains. This set is not comprehensive but does cover a wide range of site styles and topics. Multiple experimenters worked simultaneously with different combinFormation seed sets [5]. The files were then merged together after the experimenters finished.

Although the labeling of surrogates was partially subjective, experimenters were given consistent instructions and guidelines concerning what to consider as junk and

non-junk. Performance in this problem space is inherently interpretive, involving some subjectivity. Thus, giving individual experimenters a role that includes subjective interpretation seems appropriate. We believe that a set of human experts is capable of making decisions about classification that will be acceptable in most cases. The role of the experts who construct training sets is similar to that of “corpus editors” [47]. In both cases, we observe that there may be ethnographic issues in choosing a representative set of experts. These issues deserve further study, as they are relevant to training set or ontology construction in any situation in which the classification is a partially subjective.

2. Results

PCA projections were constructed for each dataset in order to obtain a visual sense of the underlying structure of the data. On each PCA scatter plot (Figures 10 - 15), blue solid crosses represent junk surrogates, while red outlined circles represent non-junk surrogates. Ellipses have been drawn to show the equiprobable contours 2 standard deviations away from the mean. This corresponds to the boundary that contains 95% of the data for each class. The center of each ellipse represents the mean of the distribution for the given class. A data sample of the appropriate class is most likely to be found near the mean.

3. The Structured Collection

Shown in Figure 10, the PCA projection for text surrogates in the Structured Collection indicates that the majority of the junk samples cluster in a more confined region of feature space than non-junk samples. The distributions for junk and non-junk appear to be unimodal. The distribution of the text data looks Gaussian, with the non-junk apparently more symmetric around the mean than the junk. For the

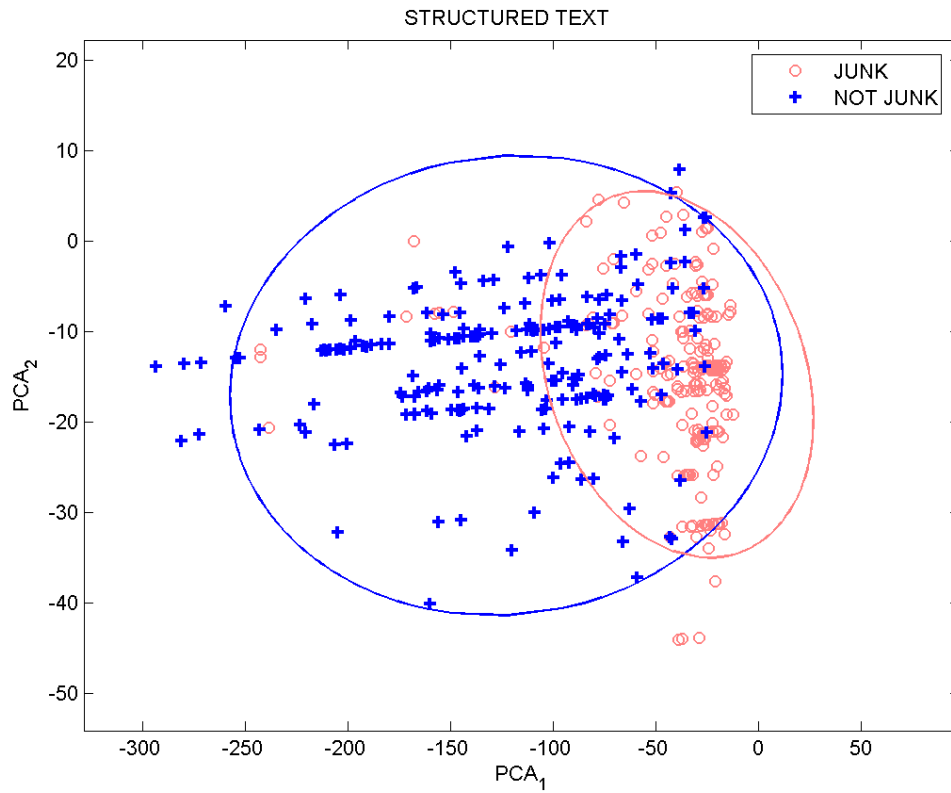


Fig. 10. 2D PCA scatter plot of text surrogate candidates in the Structured Collection.

image surrogates, the distribution looks less Gaussian, and so the assumption of the classifier might be questioned (see Figure 11). Classification performance with 5-fold cross-validation was estimated at 79% (10% standard deviation) for text surrogates, and 75% (10% s.d.) for image surrogates. The lower classification performance for image surrogates is consistent with the observation that its feature space is noticeably less Gaussian than the text surrogate space. This result suggests that better performance may be obtained with alternative classification algorithms.

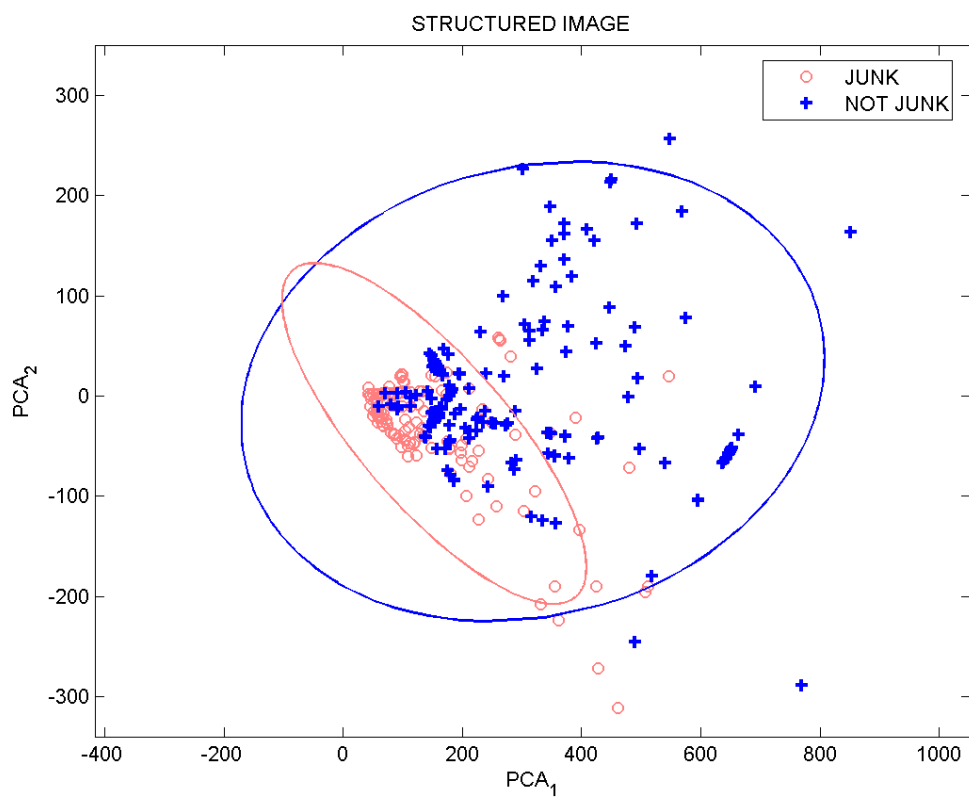


Fig. 11. 2D PCA scatter plot of image surrogate candidates in the Structured Collection.

Table II. Test data characteristics and performance results with cross validation.

Collection	Web sites	Surrogate Type	Data (number)	5-Fold Cross Validation	
				performance (%)	standard deviation
Structured	news sites, EverQuest II sites, travel sites	text	515	78.74	9.85
		image	493	74.62	10.30
Non-Structured	small web pages by Google search	text	204	82.44	4.22
		image	284	81.57	14.21
Complete	(Non-Structured + Structured) web sites	text	719	81.94	4.81
		image	777	78.30	7.12

4. The Non-Structured Collection

The PCA projections for the Non-Structured Collection (Figure 12 and Figure 13) show a very similar structure to the one for the structured pages, both for the text and image feature spaces. This suggests that the sample distributions are fairly general over a wider range of web sites. This is an encouraging sign for using one classification method over diverse sites. The classification rate is again hampered by the fact that the distribution of junk data and non-junk data is not strictly Gaussian, which violates the main assumption of the quadratic classifier. However, class separability for the Non-Structured Collections is higher than in the Structured Collection, as indicated by the 5-fold cross-validation estimates: 82% (4% s.d.) for text surrogates, and 82% (14% s.d.) for image surrogates.

5. The Complete Set

The PCA projections for the Complete Set (Figure 14 and Figure 15), are very similar to the projections for the Structured and Non-Structured Collections. As before, the data appears to be unimodal, but not strictly Gaussian. At the same time, the Complete Set appears more Gaussian than the separate individual data sets. That the

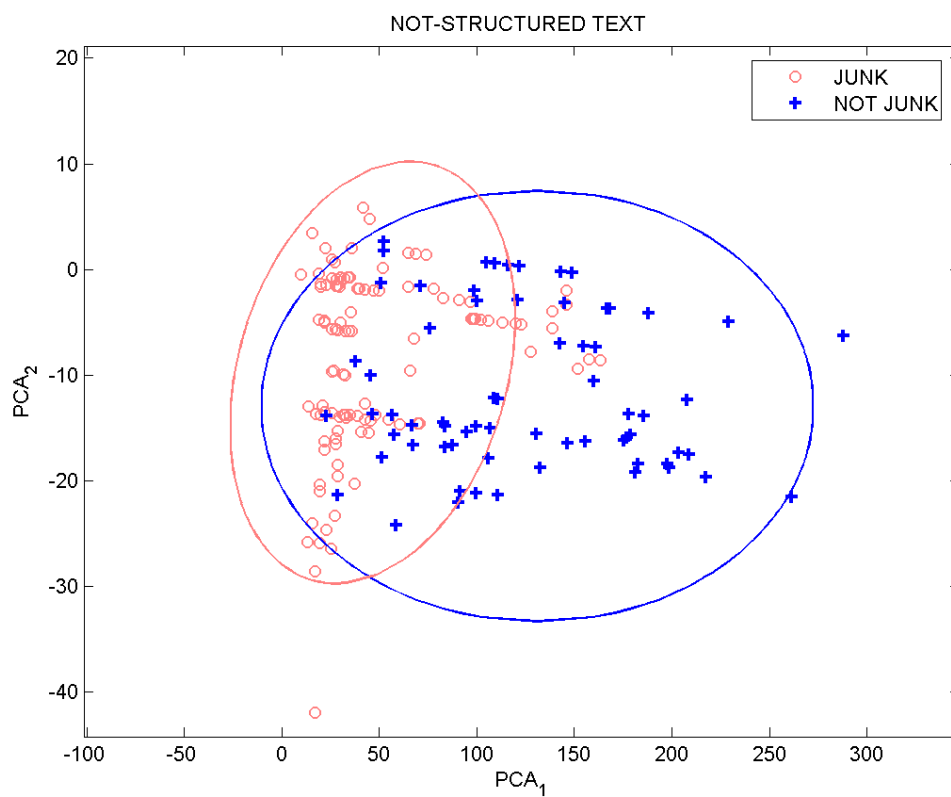


Fig. 12. 2D PCA scatter plot of text surrogate candidates in the Non-Structured Collection.

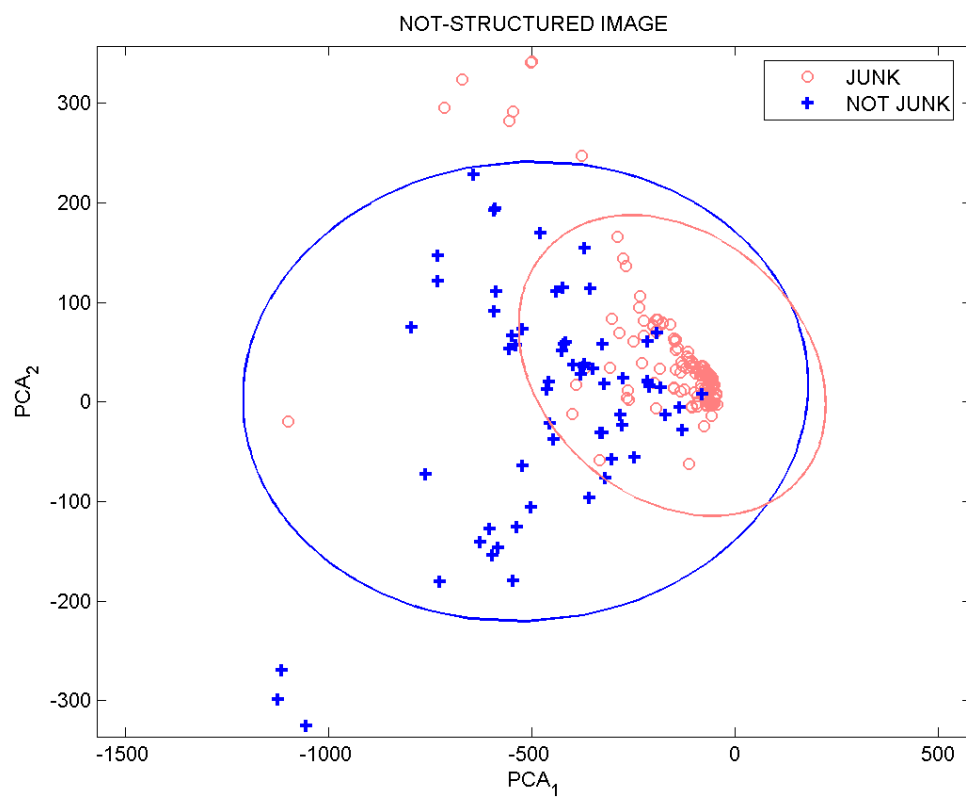


Fig. 13. 2D PCA scatter plot of image surrogate candidates.

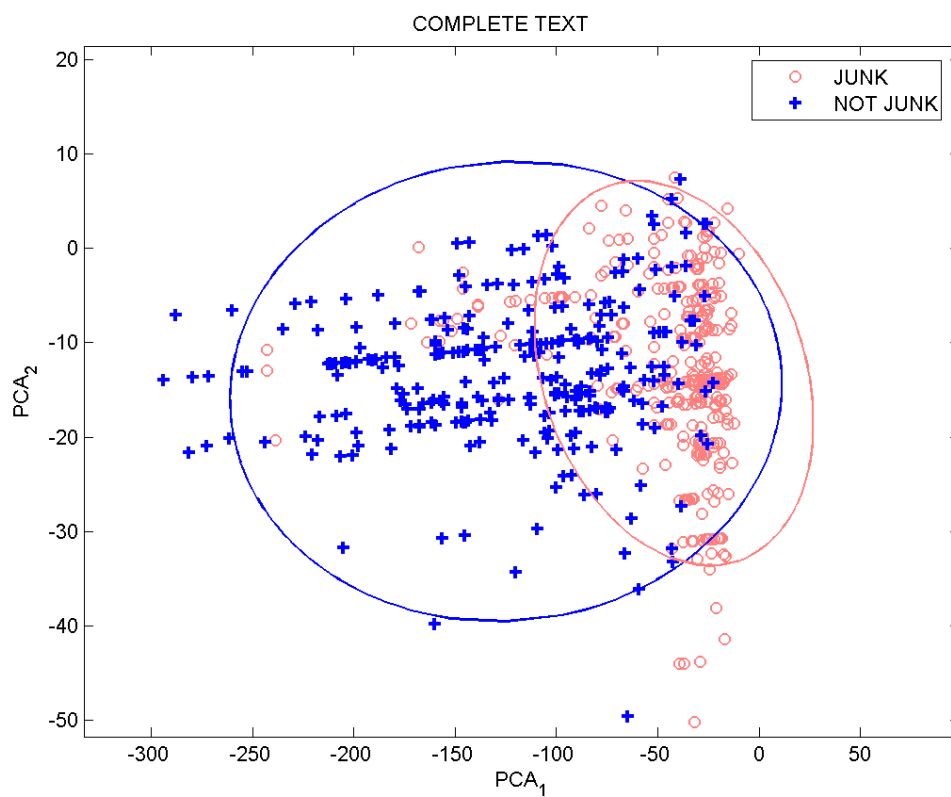


Fig. 14. 2D PCA scatter plot of text surrogate candidates in the Complete Set.

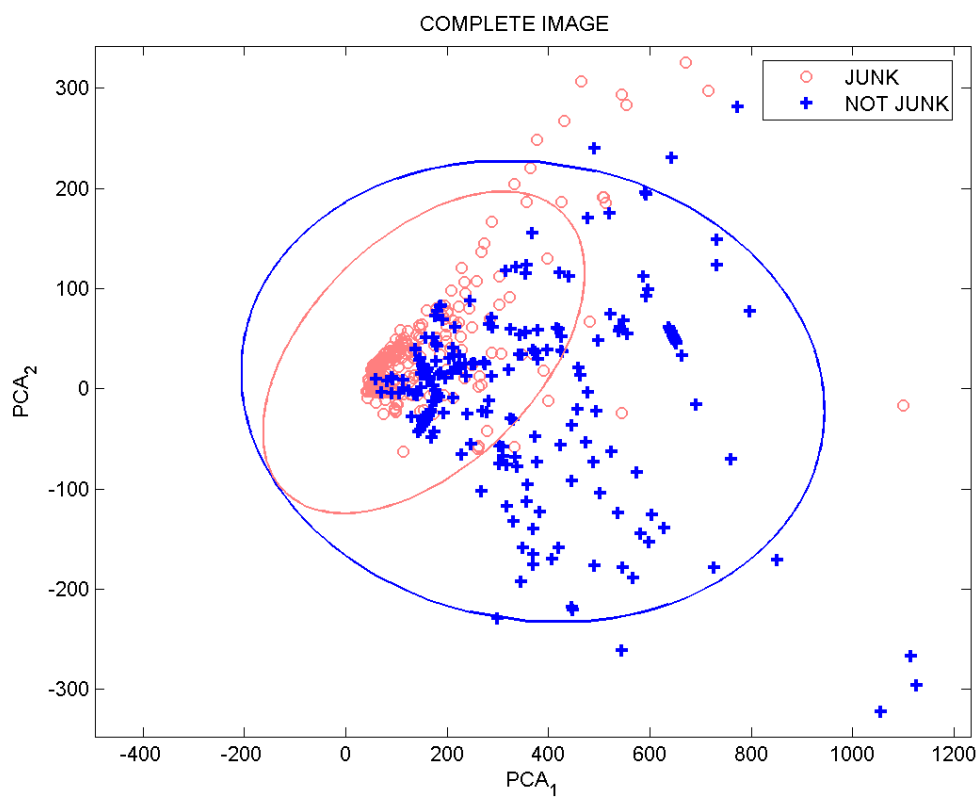


Fig. 15. 2D PCA scatter plot of image surrogate candidates in the Complete Set.

shape of the distribution grows more Gaussian as the amount and diversity of the data increases indicates that this classification method is applicable to HTML documents in general. Classification performance with 5-fold cross validation was estimated at 82% (5% s.d.) for text, and 78% (7%) for images. These average performance values are within the bounds defined by the performance on each separate collection, indicating that the quadratic classifier is able to find structure that is common to both types of web pages. It is interesting to note that the standard deviation for the Complete Set on the image feature space is half of that for either Collection, a result that may be partially explained by the fact that the Complete Set has more surrogate examples, and therefore lower variability from fold to fold. The lower variance of the Complete Set also suggests that the quadratic classifier becomes more stable when trained on a more diverse sample of web pages.

F. Discussion

Performance measures for the three datasets and feature spaces are summarized in Table II. The average performance of the quadratic classifier (80%) is significantly above chance level (50%), indicating that the selected features in Table I do contain discriminatory power about the information content of the candidate surrogates. However, the standard deviation across folds is somewhat high, particularly for the smaller Collections. We believe this is due to the presence of outliers in the data, which may have been caused by labeling errors made by the human experts.

The PCA projections and classification results indicate that there is significant discriminatory structure in the data. The question remains as to the extent to which this structure is successfully captured by a quadratic classifier, since the class densities are clearly asymmetric. Nonetheless, the utilization of a quadratic classifier is a

successful initial move towards the application of pattern recognition techniques to the problem of identifying junk surrogate candidates.

Our first algorithmic approach incorporates features of DOM into surrogate formation and classification. It forms a linear set of surrogates from the document and filters them later. This strategy does not make sufficient use of the document's powerful tree structure. Before developing the new algorithm, we had to collect and label better test collection for validation. Thus, the next chapter presents the test collection system which enables us to systematically collect and label the test data for the evaluation of the new algorithm.

CHAPTER V

TEST COLLECTION MANAGEMENT AND LABELING SYSTEM

The problem with the algorithm presented in the previous chapter was the presence of outliers in the test data. Thus, we set out to collect test data more systematically.

Many information retrieval or information extraction researchers use test collections to validate their algorithm performance with the metrics such as precision and recall computed in reference to a test collection [48]. The *test collection* consists of a set of documents, a clearly formed problem that an algorithm is supposed to provide solutions to. A test collection is *labeled*, that is, annotated with the answers that the algorithm should produce when executed on the documents. Test collections are an important factor in research validation, so they need to be built objectively and maintained consistently. There are publicly available test collections, developed by institutions such as TREC [49]. However, the documents in those collections are not labeled in a manner appropriate for all information retrieval and extraction research problems. Thus, researchers since the initiation of the Open Video test collection [50] and before have needed to build and label their own test collections.

To lighten researchers' burden of building their own test collections, we developed a *test collection management and labeling system* (TCMLS). A systematic mechanism for building test collections eliminates errors and enforces consistency in labeling practices. In addition, the system that manages test collections facilitates usability in the process of building test collections, applying them to validate algorithms, and potentially sharing them across the research community.

The system is designed in a client-server model. The client is implemented as a Firefox browser extension, which enables researchers to collect and label any HTML documents using their browser. The extension sends a service request message, such

as what document needs to be labeled. The server, which is built using the open source *Semantic Distributed Computing Services* [51], performs the requested service, such as uploading a copy of the document, and sends a response, including result status. The client and server specify the request and response message format using XML. If the message does not follow the specified syntax and semantics, the server will ignore it.

This section starts by examining related work addressing the need for managing test collections. Then, we present the design of the TCMLS, in the context of our research problem, which involves extracting informative parts from web documents. We describe how the system works, and how we are using it. We close by discussing how this work can meet the research community’s needs.

A. Background

Various test collections have been used throughout the years for the evaluation of information retrieval systems. TREC provides a set of large reference test collections that are extensively used by researchers [49]. TREC collections have included Web Test Collections, the Blog Track, the Query Track, the Question Answering Track, and the SPAM Track. It also supported a video track devoted to research in automatic segmentation, indexing, and content-based retrieval of digital video, which then emerged as independent [52]. INEX is an initiative for the evaluation of XML retrieval [53]. This initiative provides an opportunity for participants to contribute to the construction of a large test collection and to evaluate their retrieval methods using uniform scoring procedures and a forum to compare their results.

This would seem to be a large set of test collections. However, in fact, the utility of these collections for research is limited to the evaluation of an important but

small set of possible information retrieval and extraction problems. To address this limitation, many researchers build their own test collections to enable conducting performance evaluation. For example, Dakka *et al.* could not rely on the above popular research collections because the collections do not include the variety of alternative news sources in news portals, which is critical in their research [54]. Instead, they collected news articles crawled and processed by Newsblaster [55] and conducted user studies to collect their associated relevance judgments for their experiments. Liu *et al.* collected PDF documents from various sources, which they use to conduct a user study to evaluate the quality of their table detection algorithm [56]. Song *et al.*, solving an information extraction problem similar to ours, collected 600 web documents from 405 sites in 3 categories in Yahoo: news, science and shopping [57]. Both research problems involve using the Document Object Model (DOM) tree representation of an HTML document [2] to identify document components. They created their own tool for labeling importance of blocks in web documents. Five human assessors manually labeled blocks in the documents with importance values. They then used this labeled test collection to assess the precision of their extensions to the VIPS algorithm for automatically assessing block importance [58]. A disadvantage of this tool is that instead of being able to label any DOM nodes, the blocks that can be labeled are only those identified by VIPS. Thus, the tool has limited extensibility for building test collections for validating solutions to different research problems. However, the TCMLS system can label any informative blocks in documents, as well as specific metadata, such as images and image captions.

B. System Design

Building a test collection is conducted through three stages. The first stage is defining how documents will be labeled and categorized. The next is interactively collecting and labeling documents. Finally, algorithms operate on the test collection, and compare their results with the labels. This paper focuses on the processes of first two stages, which the TCMLS is being developed to support. We designed the system in a client-server model in order to manage a central test collection repository.

This section presents the semantics with which our test collection documents can be labeled. Once such semantics are defined, the test collection can be formed. TCMLS usage begins with the user identifying each document to collect, in response to which the TCMLS stores a copy of the document and its media assets in its repository. Next, the user applies the semantics to each test collection document using the TCMLS, and the system stores the labels with the associated document in the repository.

1. Document Labeling Semantics

We defined labeling semantics for the research problem of extracting informative images and text from a document. Table III describes these labeling semantics. The labels are annotated to the appropriate DOM nodes (elements) in an HTML document. We also enable labeling the document as a whole as belonging to one or another `category`. So far, in our case, the `category` labels for documents are either “news article” or “news index”. The labeling semantics are coded in the TCMLS.

Table III. Document labeling semantics for the test collection to validate informative images and text extraction algorithm.

<code>category</code>	The category the test document belongs to.
<code>partition</code>	For partitioning a document to identify semantic sub-trees of informative context. Partitions are not nested or overlapping in the DOM tree.
<code>inform_img</code>	Label informative image with an appropriate caption and the best field true for the one best representing the content.
<code>inform_text</code>	Label informative text.
<code>noninform_text</code>	Label non-informative text only within a block informative text to restrict scope of the labeling.
<code>caption</code>	Mark as a caption text that describes an informative image.

The image shows a screenshot of the BBC News website with several DOM Inspector annotations. The main headline is "Friendly fire pilot back in Iraq". A photo of Lance Corporal of Horse Matty Hull is shown with the annotation "inform_img". The caption below the photo is "Capt Matty Hull died four weeks ago in the attack in Basra" with the annotation "caption". The sub-headline "A US pilot involved in the friendly fire killing of a UK soldier is returning to fight in Iraq next month, it has emerged." has the annotation "inform_text". The main article text "Lance Corporal of Horse Matty Hull, 25, of Windsor, Berkshire, died when his Scimitar tank came under fire from a US A-10 'Tank Buster' plane in March 2003." has the annotation "noninform_text". The article continues: "One of two pilots involved in the incident is now being deployed in Iraq as part of the Idaho Air Guard. A spokesman said he was deployed due to his 'extensive combat experience'." Below this is the sub-headline "'Unlawfully killed'" and the text "Air Guard spokesman 1st Lt Tony Vincelli said the squadron would focus on providing close air support to ground troops, but for security reasons the exact location of". The annotation "inform_text" is placed over this text. On the right side, there are sections for "WHERE I LIVE" (BBC Berkshire), "SEE ALSO" (Drawing solace after a four-year quest, Search for truth on 'friendly fire' death, 'Friendly fire' widow's Bush plea, 'Friendly fire' family view video), "RELATED INTERNET LINKS" (Ministry of Defence), and "TOP BERKSHIRE STORIES" (Pub locals win on namesake horse).

Fig. 16. Example of nodes highlighted with Modified DOM Inspector and labels assigned to the test collection document.

2. Interactive Collecting and Labeling Client

The TCMLS client enables researchers to collect any HTML document using their browser. It is implemented using the *DOM Inspector* (DI) [48], an open source Firefox extension. DI enables the user to examine the hierarchical DOM tree of the HTML source code of a web document [59]. The DI already contains a built-in feature that allows the user to add, edit, or remove attributes of document elements. When you click an HTML element in the DI, the corresponding document block is highlighted with a red rectangle box in the browser (see Figure 16). Details about the HTML element, in the form of attribute value pairs, are displayed. We extended this software by creating the aptly named *Modified DOM Inspector* (MDI), which enables the user to easily label appropriate nodes in test collection documents.

3. Identify Each Document to Collect

Figure 17 shows the buttons that we added to the MDI to enable the user to apply label semantics to a DOM element. The researcher uses Firefox to browse. She selects a document to collect, and assigns a `category` from the set offered by the Modified DOM Inspector. She clicks the URL to Server button (see Figure 17). Then, the MDI extension forms the `collect_document` message, in XML, which requests the system server to store a document in the test collection repository, as follows:

```
<collect_document category="" url="" datetime=""/>
```

In the `collect_document` message, the `category` field specifies the category selected with the interface. The `url` field is for the document URL that the researcher is collecting, and the `datetime` is the time that the message is sent to the server. This step, and the subsequent performance of the service to store the document in the test collection repository, must be performed prior to interactive labeling.

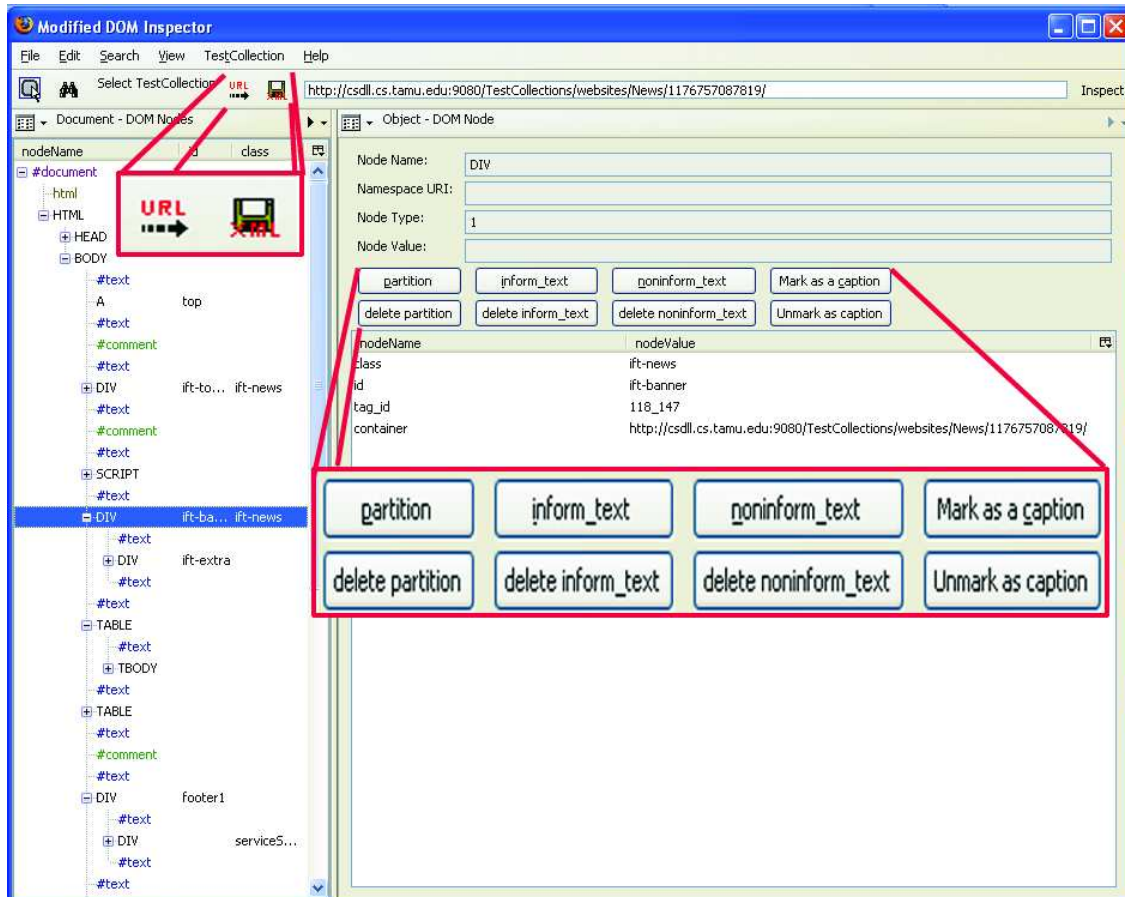


Fig. 17. Modified DOM Inspector: URL to Server button stores document in repository. Semantic buttons (right) label selected HTML element. Save XML button stores labeling in repository.

<http://csdll.cs.tamu.edu:9080/TestCollections/websites/News/1176757087819/index.html>

BBC NEWS Home News Sport Radio TV Weather Languages

UK version International version | About the versions Low graphics | Accessibility help

News services
Your news when you want it

News Front Page
Last Updated: Sunday, 15 April 2007, 14:46 GMT 15:46 UK
E-mail this to a friend Printable version

Friendly fire pilot back in Iraq
A US pilot involved in the friendly fire killing of a UK soldier is returning to fight in Iraq next month, it has emerged.

Lance Corporal of Horse Matty Hull, 25, of Windsor, Berkshire, died when his Scimitar tank came under fire from a US A-10

WHERE I LIVE
BBC Berkshire
Sport, entertainment and more from the BBC Berkshire website

SEE ALSO
Drawing solace after a four-year quest
16 Mar 07 | UK
Search for truth on 'friendly fire'

RELATED INTERNET LINKS
Ministry of Defence
The BBC is not responsible for the content of external internet sites

TOP BERKSHIRE STORIES
Pub locals win on namesake horse
Trains on track after 10-day halt
Community centre plans move ahead

'Unlawfully killed'
Air Guard spokesman 1st Lt Tony Vincelli said the pilot's squadron would focus on providing close air support for ground troops, but for security reasons the exact location of the deployment would not be made public.

The other pilot involved in the "blue on blue" attack on British

http://newsimg.bbc.co.uk/media/images/42687000/jpg/_42687225_matty_pa203b.jpg

```
eunyeetwang:/project/ecologylab/TestCollections/News/1176757087819$ ls -a1F
total 80
drwxrwxr-x 7 eunyeec ecology 4096 2007-04-16 16:02 ./
drwxrwxr-x 289 eunyeec ecology 8192 2007-10-16 11:58 ../
-rwxrwxr-x 1 eunyeec ecology 41180 2007-04-16 15:59 index.html*
-rwxrwxr-x 1 eunyeec ecology 1984 2007-04-16 16:02 label.xml*
drwxrwxr-x 8 eunyeec ecology 4096 2007-04-16 15:59 news.bbc.co.uk/
drwxrwxr-x 6 eunyeec ecology 4096 2007-04-16 15:59 newsimg.bbc.co.uk/
drwxrwxr-x 3 eunyeec ecology 4096 2007-04-16 15:58 newsrss.bbc.co.uk/
drwxrwxr-x 2 eunyeec ecology 4096 2007-04-16 15:59 stats.bbc.co.uk/
drwxrwxr-x 4 eunyeec ecology 4096 2007-04-16 15:58 www.bbc.co.uk/
```

Fig. 18. Above: an example test web page (index.html) and associated resource image (_42687225_matty_pa203b.jpg) stored in the repository; Below: the directory structure stored for the Above test web page in the repository.

4. Store the Document in Repository

When the server receives the `collect_document` message, it performs the service by connecting to the specified URL, retrieving the document, and storing it and referenced resources such as images and JavaScript in the repository. As the documents themselves and the associated resources can be changed or removed, we stored copies. This requires resolving all URLs for referenced resources into relative paths stored in the repository. Hyperlinked documents were not stored and links to them were not transformed. The TCMLS stores and fixes resource references in order to preserve complete visual copies of each document (see Figure 18).

Some HTML documents do not follow the specification completely because even though there are some missing ending tags, the browser renders those documents without any problem. Thus, a typical XML parser is unable to form the DOM tree from them. To address this issue, we use JTidy [60], a syntax checker and pretty printer. JTidy cleans up malformed and faulty HTML, so that the TCMLS can build DOM trees from any document in the test collection repository.

Upon forming the DOM for an HTML document, the system also generates a unique identification number, `tag_id`, for each HTML element in test documents. This enables the cross-reference between the label and the test document in separate files. The generated `tag_id` was added as an attribute in each element in test documents in the repository. The following is the example from a test document.

```
<html tag_id="0_1144" lang="en">  
  
<head tag_id="1_39">
```

We used the depth-first-search (DFS) algorithm to generate `tag_id`. The DFS algorithm records the discovery time and the finishing time as each element in the DOM

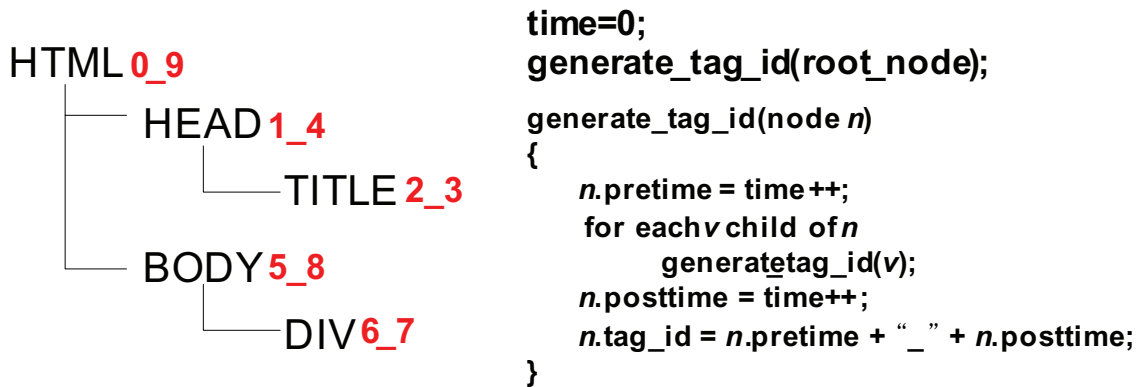


Fig. 19. Generate `tag_id` of each HTML tag by traversing the DOM tree with DFS algorithm. The left diagram shows the DOM tree, with each `tag_id` generated by algorithm (right).

tree is traversed. We defined the `tag_id` by combining the discovery time and finishing time. The `tag_id` is unique in the DOM tree, and also provides parent-child relationship among the elements in the DOM by inspection of the number range in the `tag_id`.

For example, in Figure 19 left, document elements have been labeled with `tag_id`. The `tag_id` of the HTML is 0_9, the HEAD is 1_4, and the BODY is 5_8. The starting and ending range of the `tag_id` shows that the HTML element is the parent of both the TITLE and the HEAD, and that the TITLE is in the different tree from the BODY. Knowing the parent-child relationship helps researchers to label the test documents without having redundant labels inside the same sub-tree. It also helps to locate the labeled tags in the test documents. This is a more efficient way to label DOM node relationships than XPath, which has functionalities to find relative nodes, because XPath incurs tree traversal iterations to operate, while our `tag_id` directly represents parent/child relationships.

After the system server processes the request to store the document in the repository, it sends the client a response. The response is either `ok_response` (the request

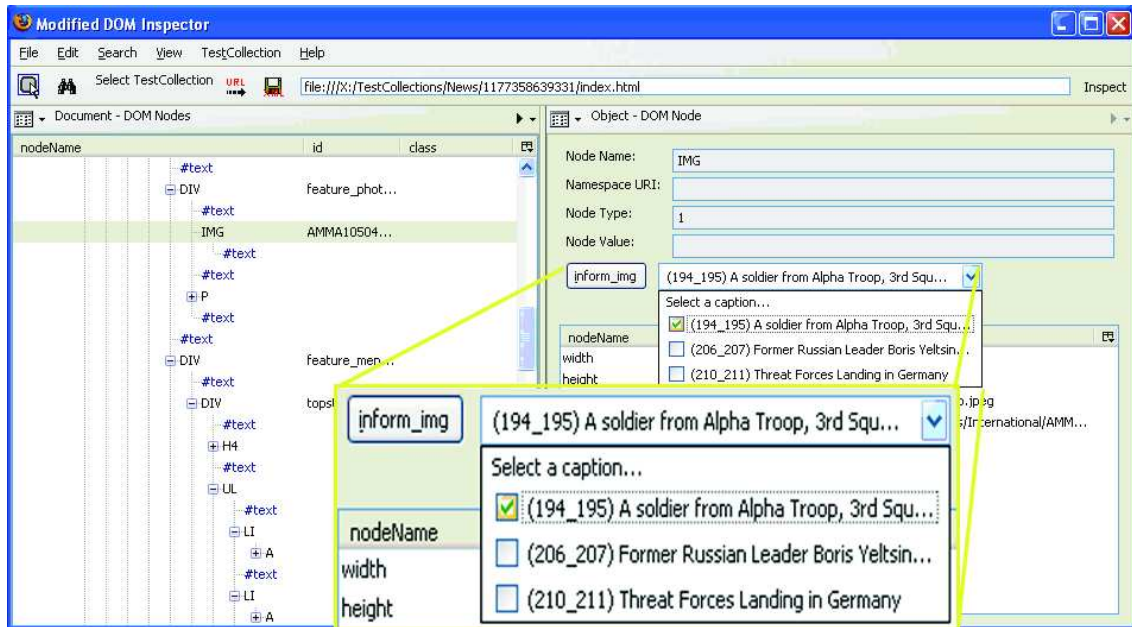


Fig. 20. Interface for labeling the informative image with its caption in the Modified DOM Inspector.

has been successfully finished) or error_response (the request failed to be performed by the server). When the client receives the ok_response, the browser redirects to the test document URL stored in the repository, so that researchers can label it.

5. Label Each Document

The MDI is used to label each document, with the semantics described in Table III, by clicking labeling buttons (see Figure 17, right). When an img element is selected, another view appears (see Figure 20). This view enables labeling each image as informative or not, and, in the former case, for one or more captions to be associated. A dropdown checklist will contain all captions that have been labeled through the interface in Figure 20. The user can check and uncheck captions to associate them with the correct image.

6. Store Labels in Repository

When the user has finished labeling a document, she clicks the Save XML button (see Figure 17). Then, a function recursively walks through the HTML DOM tree with the DFS algorithm, and checks for the annotation of the labels. Each label is represented with its `tag_id` to form compact XML that represents the labeling. Post-processing is performed to clean up the XML, such that `partition` labels are ordered from least to greatest, and each `inform_img` is associated with the correct `caption`. The completed XML labeling string is sent to the server to store in the repository. Here is an example:

```
<document
url="http://csdll.cs.tamu.edu:9080/TestCollections/websites/
News/1176757087819/" title="BBC NEWS | UK | England | Berkshire |
Friendly fire pilot back in Iraq">
  <partition_set>
    <partition id="0" tag_id="362_700">
      <noninform_text_set>
        <noninform_text tag_id="428_433"/>
        <noninform_text tag_id="434_449"/>
      </noninform_text_set>
      <inform_text_set>
        <inform_text tag_id="366_367"/>
        <inform_text tag_id="372_451"/>
      </inform_text_set>
      <inform_img_set>
        <inform_img tag_id="379"
```

```

url="newsimg.bbc.co.uk/media/images/42687000/jpg/_42687225
_matty_pa203b.jpg" best="true">
  <caption_set>
    <caption tag_id="380_381" value="L/Cpl Matty Hull died
    four years ago in the attack in Basra"/>
  </caption_set>
</inform_img>
</inform_img_set>
</partition>
</partition_set>
</document>

```

The `label_document` message encapsulates the XML labeling string, to send it to the server for storage in the repository:

```

<label_document>

XML labeling string

</label_document>

```

C. Building the Test Collection

All the built test collections can be browsed from <http://csdl1.cs.tamu.edu:9080/TestCollections/websites/>. The directory structure is based on the selected category of test documents. If a user clicks a **category**, all the collected documents under the **category** are listed. Users can easily browse and download the test documents with the label XML files.

D. Discussion

We have developed a system to reduce researchers' tedious task of test collection management and labeling. With minimal training, two undergraduate students and one graduate student were able to use the system to collect and label more than 500 documents. The collected documents were further utilized in the evaluation of the algorithm presented in the next chapter.

The system provides usability for iteratively building test collections. It facilitates algorithm validation. As they are developed, test collections are published on the web, enabling sharing by the research community. By installing the browser extension on Firefox, other researchers can also contribute to the test collection. They can browse and download the built collection, and use it for the algorithm validation. Our goal is to maintain our system to enable sharing and extending collections among the research community, to support algorithm development efforts.

While institutionalized test collections have been developed to promote solutions to important research problems, there is a world of important research problems they have not addressed. However, test collections are necessary for much research on information retrieval and extraction. The burden of creating test collections may function as a barrier to entry for important new research areas. The present research develops tools to support test collection management and labeling. It thus has the potential to facilitate the diversification of research efforts in the fields of information retrieval and extraction, by reducing the efforts necessary to address research problems whose significance has not yet been institutionally acknowledged, but which may turn out to be of great importance. The algorithm presented in the next chapter addresses the research problem of forming visual representations for documents by automatically extracting informative content. As the research area is new and the

practitioners of similar research have not publicly provided test collections, we built our own test collection with this test collection system to evaluate our algorithm.

CHAPTER VI

INFORMATION EXTRACTION ALGORITHM: EXTRACTING IMAGE+TEXT
SURROGATES VIA RECOGNIZING INFORMATIVE CONTENT FROM WEB
PAGES

This chapter presents an unsupervised algorithm that constructs an image+text representations of documents. The goal of this algorithm is an automatic extraction of informative contents from most web pages to form a visual representation, image+text surrogates. In addition, we develop the algorithm to be performed as a component in human centered interactive computing systems, so that when people search for specific information, the algorithm can rapidly process the relevant documents and generate image+text surrogates just in time for the user. Thus, we can apply our algorithm not only in servers that index documents and extract surrogates, but also in interactive systems to generate surrogates on the fly.

This algorithm improves upon our initial approach presented in Chapter IV by eliminating dependency on training data and utilizing some of the discriminatory features that contributed most in the classification. Our initial approach formed surrogate candidates from any sub-tree in the DOM, and filters the non-informative surrogates later. Instead, the new algorithm ranks the sub-trees, and determines which of them is most important from the ranking, and then forms image+text surrogates only from this sub-tree.

In order for the algorithm to work in real world contexts, we break the problem down into three stages (Figure 21). We begin with the observation that some pages on the web function as *informative content pages* (Figure 22 bottom), while others, each of which consists essentially of a set of links, function as *index pages* (Figure 22 top). Thus, stage 1 of the algorithm recognizes whether a page is a content page or

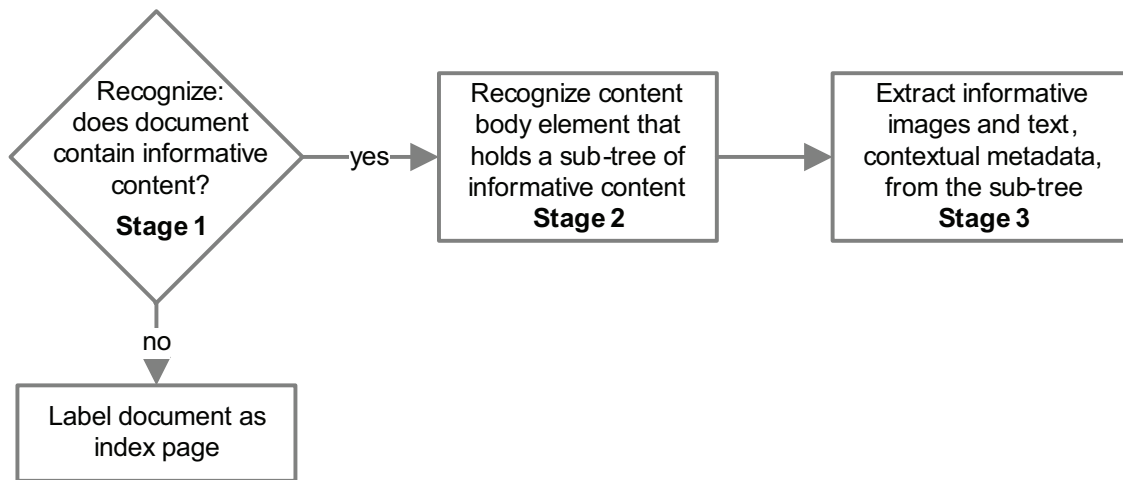


Fig. 21. Three stages of our information extraction algorithm. Stage 1 determines the page categorization, Stage 2 recognizes the informative sub-tree of the content body page, and Stage 3 extracts representative images and text from the sub-tree.

an index page. Moving forward, we observe that on the web, even within content pages we find some areas that function primarily as navigation, while others are truly informative. Thus, only for content pages, stage 2 of the algorithm identifies the most *informative content body* within a content page (rectangle border in Figure 22 bottom), and discards the rest. Once this informative content body is identified, we then set out, in stage 3, to form integrated image+text surrogates from within it.

This three stage method for extracting informative images and text from documents can be utilized in search engines and digital libraries to automatically form representations of documents to present to users. Further, the output from stage 2 of the present algorithm, which recognizes the informative parts of documents, can be applied in other information extraction and retrieval problems and algorithms, as a pre-processing stage, to improve their performance.

Web: CNN News CNN Videos

CNN.com

HOME WORLD U.S. POLITICS CRIME ENTERTAINMENT HEALTH TECH TRAVEL LIVING BUSINESS SPORTS TIME.COM VIDEO REPORT IMPACT

Hot Topics » Your Money » Iraq War » Election Center » more topics »

Weather Forecast Editors: International Set Prof

updated 12:47 a.m. EST, Wed February 27, 2008

Make CNN Your Home Page

Castro's brother faces big challenges in Cuba

Cuban President Raul Castro is taking over leadership of a country whose government believes its citizens are not working hard enough. The new president, who took the reins of power Sunday from his ailing brother, Fidel, 81, has said the country must become more productive. [full story](#)

Meet the superdelegates

Will they vote like you? Find out who your state's superdelegates are

Cooking with alcohol 101

Let your liver "on your terms"? Spare alcohol can be put to good use in these recipes

Take action

When disaster strikes or horrible events unfold, there are opportunities to effect change

Latest News

- Clinton, Obama spar over campaign tactics 14 min
- Magnitude 4.7 earthquake rattles UK
- Yousif rubs face with hands, says 'no hurt'
- Failed switch, fire cause Florida outage
- Obama tops new national polls
- Ticker: Clinton hits Obama on Farrakhan remark
- Martin: A spiritual graduation
- CNNMoney: Foreclosure bill faces veto
- Long-lost photo of Anne Frank's 'true love'
- Dr. Gupta: When to fly, when to stay grounded
- Jacko's 'Neverland' may be auctioned
- Police release tape of killer's voice
- Dad chases down 4-year-old's kidnapper
- The magic number for getting a loan
- Teacher, on tape, rails at 'stupid' 5-year-olds
- CNN Wire: Latest updates on top stories

all news from the past 24hrs »

Popular News

- Army general: 'No reason to doubt' Obama's story
- Shark tour leader was warned of danger, diver says
- Too pretty to fly?

all most popular »

Video

- Dad chases girl's kidnapper 1:56
- A stab at politics 2:22
- Teacher caught yelling at kids 1:37

more video »

What's Your 2008 Credit Score?

Excellent	750 - 850
Good	660 - 749
Fair	620 - 659
Poor	350 - 619
I Don't Know	????

[Find out instantly!](#)

CNN TV

Programs

Prime-time Exclusive!

Janel Jackson joins Larry for a rare, live, sit-down interview. The actress and singer gets personal about her love life, weight, and new album! Thursday, 9 p.m. ET

'American Morning' Weekdays, beginning 6 a.m. ET

CNN TV Schedule | Headline News | Listen to CNN TV

Scottrade

HELPING YOU RETIRE THE WAY YOU WANT.

Open a No-Fee IRA

- Powerful Research Tools
- All Listed ETFs
- Over 10,000 Mutual Funds

CNNMoney.com

A Service of CNN, Fortune & Money

Symbol Get Quote Keyword Search

- Subscribe to Money
- Free Trial
- Magazine Customer Service

Home Business News Markets Personal Finance Real Estate Technology Small Business Luxury Fortune My Portfolio CNN.com

White House to veto foreclosure bill

\$4 billion housing bill is too expensive for the administration and would 'slow the recovery of the housing sector.'

February 26 2008: 5:25 PM EST

WASHINGTON (AP) — The White House promised on Tuesday to veto a bill seeking to follow up the recent economic stimulus package with several proposals to shore up the struggling housing market and reduce foreclosures.

Senate Democrats had hoped to begin debate on the housing bill on Tuesday but action has been put off until later in the week, if not later, as Republicans kept the subject on Iraq.

The Democratic housing bill would change bankruptcy laws to allow judges to cut interest rates and reduce what's owed on troubled borrowers' mortgages, provide \$4 billion to communities to purchase and rehabilitate foreclosed homes, and improve disclosure of subprime mortgage loans in hopes that borrowers won't be surprised by big payment increases.

But the White House said the \$4 billion for purchases of foreclosed homes is too expensive and "would constitute a bailout for lenders and speculators, while doing little to help struggling homeowners."

The provision rewriting the bankruptcy code, the White House said, would allow borrowers to effectively rewrite their mortgage contracts, leading lenders to tighten their standards and raise interest rates.

The White House said both provisions would in fact slow the recovery of the housing sector.

The Democratic measure also contains provisions stripped from the Senate's version of the stimulus bill to boost mortgage revenue bonds and add flexibility to help homeowners refinance subprime loans and to allow homebuilders and other money-losing businesses to reclaim taxes previously paid.

The bankruptcy measure, a similar version of which has cleared a House committee, is fiercely opposed by

Video

How one Colorado family survived the scary and challenging experience of almost losing their home. [Play video](#)

Sponsored Links

Earn a Bachelor's in 12 Mo.
Earn an BA/BS online in 12 months!
Fully Accredited. Get free info now
[www.EarnFastDegree.com](#)

Lionbridge - Outsourcing
Make Your Global Operations Successful. Download Free Whitepaper.
[www.lionbridge.com](#)

Top Stories

- Bernanke vs. the economy
- Stocks rally despite bad news
- White House to veto foreclosure bill
- Lawmakers mull \$15B mortgage bailout
- Home prices plunge

Roll over your old 401(k) to a Fidelity IRA and our Rollover Specialists will help with the paperwork.

ACT NOW

Fidelity Brokerage Services, Member NYSE, SIPC 482090.2 Smart move.™

Photo Galleries

WINNER TAKES ALL

Flight Reading

With only about 74% of flights arriving on time, it's a good idea to have a book in your carry-on. Here are three of our top choices for March. [\(more\)](#)

Smart For Two smart for few

Daimler's tiny city car turns heads, but it has more than its share of problems. [\(more\)](#)

5 smart tax moves

Fig. 22. Top: Index page; Bottom: Content page with most informative content body selected.

A. Background

Extensive prior research addresses the problem of information extraction from web pages. Detailed background about information extraction research is covered in the survey papers such as [61, 62]. We focus on methods most relevant to the present research.

Some research has applied visual features in information extraction algorithms. The VIPS algorithm is based on visual layout [58]. Chen *et al.* also used the visual information of pages rendered in the system [63]. A notable aspect of this research is that it depends on using Microsoft Internet Explorer to perform the layout of web pages, as an intermediate processing stage. A problem with this visual approach is that it requires downloading images and rendering pages to extract features for the algorithm; because this relies on networks downloads, this can be resource-intensive for interactive systems. One end product of this work is the identification of significant blocks in index pages. We are solving a different problem, which is to recognize the most informative content body of a content page document, and then to extract informative image+text surrogates from this block. Applying the visual features with our approach may be beneficial but the present research establishes that it is not essential. It is more significant for us to solve this extraction problem for interactive systems, which must operate with fewer resources on diverse platforms.

B. Information Extraction Algorithm

Often, the algorithms and technologies are designed and implemented without fully taking into account human cognitive abilities, the ways we perceive and handle information. That is, researchers and engineers often develop computing technologies in relative isolation [64]. The goal of our information extraction algorithm presented

in this section is based on knowledge of how people understand information with visual cognition. Thus, the algorithm is developed to recognize informative images and text from web pages to form better representations. To handle the variety of content presented in practice, we developed the three stages of procedure in Figure 21. We starts from recognize the page categorization.

1. Categorize Index Page or Content Page

In order to categorize pages, we investigated the primary differences, which is that only the content page contains the informative content body. Here are the main differences between the content body and the other parts of pages:

1. The content body has a greater number of significant words than the other parts.
2. The content body has very few words surrounded by the `<a>` tag compared to the other parts. The words surrounded by the `<a>` tag mostly represent the hyperlinked documents not the current document, so those words are not the significant text for the current document.
3. The ratio of the stopword count to the total word count in the content body is minimized in comparison to the other parts. The *stopwords* are words that are very frequent, and do not carry meaning, such as ‘the’. We included web stopwords such as ‘email’ or ‘advertisement’ in the stopwords list.

We developed DOM node ranking metrics that are designed to assign high weights to DOM nodes that have the above content body characteristics. As the metrics utilize the text in the node, the nodes that do not hold any text will be ignored. Here is the DOM node ranking metric:

$$\begin{aligned}
 S_{\text{node}} &= nw(\text{node.text}) - nw(\text{node.atext}) \\
 \text{rank}(\text{node}) &= \underbrace{S_{\text{node}}}_{(1)} \times \underbrace{\left[\frac{S_{\text{node}}}{nw(\text{node.text})} \right]}_{(2)} \times \underbrace{\left[\frac{nw(\text{node.text}) - nstopw(\text{node.text})}{nw(\text{node.text})} \right]}_{(3)}
 \end{aligned}$$

$nw(\text{text})$ = word counts of the text

$nstopw(\text{text})$ = stopword counts of the text

node.text = text that the DOM node is holding

node.atext = text that is surrounded by `<a>` tag and the DOM node is holding

The first parameter among the ranking metrics is (1), S_{node} , the significance. The S_{node} shows how significant the node is for the current page based on the informative textual content that the node is holding. The second parameter (2) indicates the ratio of the significant words to the all the words, and the third parameter (3) shows the ratio of non-stopwords to all the words. These parameters align with the characteristics of content bodies, so the higher the number of significant words, the ratio of the significant words, and the ratio of the non-stop words are then the higher the rank of the node will be and the probability of the node to belong in content body will become higher.

We created a data structure that maintains a certain number of highest rank nodes in sorted order. While it is parsing a page and building the DOM tree, the algorithm discovers DOM nodes and calculates ranking weights. The data structure is filled by these nodes of which the ranks are larger than 0. The nodes for which the rank is 0 contain text that is surrounded by links or stopwords. This sorted data structure becomes updated with the newly discovered node by comparing with the

lowest node in it. Thus, after the parsing is finished, the data structure will contain the highest rank nodes present in the DOM tree.

Then, the algorithm iterates through the data structure to find a common parent or grandparent node that holds these highest ranked nodes. If the algorithm cannot find a common parent or grandparent node for these highest rank nodes, it recognizes this page as an index page. Otherwise, the algorithm recognizes the page as a content page, and the common parent node is identified as the content body node.

Instead of breaking down the document from the root DOM node, we took a bottom-up approach from the highest ranked nodes to determine the content body. If we took a top-down approach, we would need to calculate and compare the ranks among different combinations of sub-trees in the DOM to find the fine-grain content body. The bottom-up approach reduces the operation's computational complexity by maintaining the sorted data structure for the highest ranked nodes. Thus, the algorithm determines the categorization of the page while it is parsing the HTML page and building the DOM tree.

Algorithm 1 presents the procedure for categorization of an index page or a content page and the diagram for the procedure is in Figure 23.

2. Recognize Informative Contents

The content body node determined by the previous algorithm holds the informative text, but it does not necessarily hold the informative images. It mostly depends on how pages are authored, but the common characteristic is that the informative images reside in the same branch from the root DOM node and are close to the informative text. The informative images reside in the context of where the informative text is, and their closeness in the document structure shows the meaningful relationship between images and text.

Algorithm 1 Categorization of an index page or a content page

Require: an HTML page

Ensure: page category (either an index page or a content page)

- 1: **SortedList** highRankNodes[k]; (sorted data structure, and k is an arbitrary number which is set to 10 in the current implementation)
 - 2: **Node** contentBody;
 - 3: **while** building a DOM **do**
 - 4: Maintain the highRankNodes with the highest ranked nodes;
 - 5: **end while**
 - 6: **while** iterate highRankNodes **do**
 - 7: Check the common parent node of the each entry;
 - 8: **end while**
 - 9: contentBody = the common parent or grandparent node that holds most of the nodes in highRankNodes;
 - 10: **if** contentBody.exist() **then**
 - 11: **return** “content page”;
 - 12: **else**
 - 13: **return** “index page”;
 - 14: **end if**
-

Thus, we recognize the parent of the content body node as the root of the sub-tree that holds informative content. The sub-tree is presented in Figure 23, (3) with the triangle region.

3. Extract Representative Images and Text

After we determine the sub-tree that holds informative content, we select the informative images and text within it. Even in the informative sub-tree, there are non-informative images such as copyright or icon images. Thus, in the selection process, we utilize the characteristics of informative images:

- Representative images reside in the sub-tree of the nodes that hold informative contents.
- Image Size: Not too small (width is lower than 24 and height is lower than the 35). Small images are usually icons or copyright images. Not high aspect ratio (larger than 0.9). Navigation bars or advertisement images mostly have a high aspect ratio.
- Texts in the image URL or alt attribute: Eliminate images that have ‘adv’, ‘ad’, or ‘advertisement’ words.

In the image node, the size of the images may be defined in an HTML attribute, but this markup may be missing. In such cases, we need to download the image to determine the size of the image. We were trying to decide whether or not to use images that have hyperlinks as a feature, but some sites show a bigger size of the representative image with a hyperlink. Thus, we didn’t use it as our feature to recognize the representative images. However, the most other cases, images that have hyperlinks are highly probable to represent the hyperlink document not the

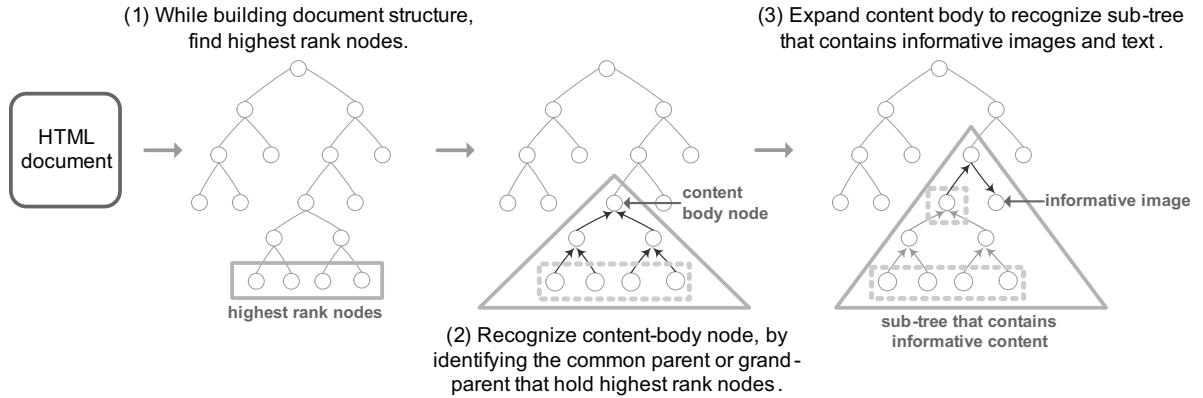


Fig. 23. The procedure of how the algorithm recognizes the content body node by identifying the highest rank nodes in the DOM.

current document. Future research will use mime type of the hyperlink destination to incorporate this feature.

Representative text is clipped from the informative texts that reside in the sub-tree of the determined content body node. We can also extract a caption text by finding the nearest text from the informative image in the DOM tree. We tried to select the text surrogate that is similar to the caption text so that we can integrate these as combined image+text surrogates.

C. Experiments

In this section, we report empirical results obtained by applying our proposed information extraction algorithm to determine the categorization of web pages, identify the content body of the pages, and recognize the representative images and text inside the sub-tree of the content body. We present the performance result of the three parts of our algorithm separately, and discuss the overall results. The results show that our algorithm achieves significant performance in finding informative images and text from documents by leveraging the DOM structure and semantic features associated

with images and text.

1. Dataset

We collected a test data set for our experiment. So far, we collected 239 article pages from news sites, which are 80 pages from CNN, 52 from the BBC, 54 from ABC, and 53 from Scientific American. We also collected index pages from same sites. They are 27 pages from CNN, 77 pages from BBC, 23 from ABC, 36 from NYTimes, and 15 pages from Scientific American. We call this test data set as news collection. We created another test data set calls research collection. In this collection, we collected pages about research from university labs and from research center sites such NSF, PARC, Microsoft Research, IBM Research, and Los Alamos Laboratory. We collected 151 research article pages and 103 research index pages.

The test data set is publicly available for sharing purposes at:

<http://csdll.cs.tamu.edu:9080/TestCollections/websites/>

2. Evaluation Metrics

In a statistical classification task, the *precision* for a class is the number of true positive (i.e. the number of items correctly labeled as belonging to the class) divided by the total number of elements labeled as belonging to the class (the sum of true positives and false positives). *Recall* is defined as the number of true positives divided by the total number of elements that actually belong to the class (i.e. the sum of true positives and false negatives). *F-measure* is the weighted harmonic mean of precision and recall. The traditional F-measure, known as *F1*, is the evenly weighted mean of precision and recall. The accuracy is the ratio of the number of all correctly labeled to the total number of elements.

Table IV. Contingency table with the news collection: the experimental results of the page categorization algorithm. ‘Correct’ means the correct category of the pages, and ‘Assigned’ means the category assigned by the algorithm.

	Correct='Content'	Correct='Index'
Assigned='Content'	275	25
Assigned='Index'	2	152
Total	277	177

Table V. Performance of page categorization algorithm with the news collection.

Precision	0.917
Recall	0.993
F1	0.953
Accuracy	0.941
Error	0.059

3. Experimental Results: Page Categorization

In the news collection, we had 277 content pages, and 177 index pages. The algorithm recognized 275 pages correctly as content pages, while miscategorizing 2 pages. It also recognized 152 pages correctly as index pages, and failed to categorize the remaining 25 pages (see Table IV). Table V shows the statistical analysis of the results in Table IV. The precision is 0.917, recall is 0.993, and the F1 is 0.953. The results demonstrate the effectiveness of the algorithm.

In the research collection, we had 151 content pages, and 103 index pages. The algorithm determined 146 pages correctly as content pages and missed 5 pages. The 63 index pages are correctly categorized as index and the 40 index pages are categorized incorrectly (Table VI). Thus, the precision of the algorithm with the research

Table VI. Contingency table with the research collection: the experimental results of the page categorization algorithm. ‘Correct’ means the correct category of the pages, and ‘Assigned’ means the category assigned by the algorithm.

	Correct='Content'	Correct='Index'
Assigned='Content'	146	40
Assigned='Index'	5	63
Total	151	103

Table VII. Performance of page categorization algorithm with the research collection.

Precision	0.785
Recall	0.967
F1	0.866
Accuracy	0.823
Error	0.177

collection is 0.785, recall is 0.967, and F1 is 0.866. The analysis result is in Table VII.

We integrated the news collection and research collection to analyze the performance of the total collection. With the total collection, the precision is 0.866, the recall is 0.984, and the F1 is 0.921 (see Table VIII, IX).

We investigated the reason of low precision compared to recall by investigating failed index pages in categorization. The reason is that these index pages contain not only links but also substantial informative contents. One example of these pages is in Figure 24. Small rectangle boxes in Figure 24 are highlighting the informative content text, so the nodes holding the text will be ranked high. The parent node of those nodes is holding all these node, so the algorithm will determine the parent node as a content body. As this page has a content body, the categorization of the algorithm will be identified as a content page.

Table VIII. Contingency table with total test collection: the experimental results of the page categorization algorithm. ‘Correct’ means the correct category of the pages, and ‘Assigned’ means the category assigned by the algorithm.

	Correct='Content'	Correct='Index'
Assigned='Content'	421	65
Assigned='Index'	7	215
Total	428	280

Table IX. Performance of page categorization algorithm with total test collection.

Precision	0.866
Recall	0.984
F1	0.921
Accuracy	0.898
Error	0.102

One way to resolve this failure is to extract informative elements from the pages because the failed pages do contain the representative images and text, or we will be able to refine our algorithm and metrics to determine these pages as index pages. One possibility is that after the algorithm determines the content body, rank the content body node to determine whether it is the true content page or not by checking the link threshold.

4. Experimental Results: Informative Content Body Detection

In the news collections, 237 pages are labeled among the 277 content pages for use in our algorithm evaluation. The label contains where the content body is and what are the informative images and text. Among the 237 labeled content pages, the content body nodes of 203 pages are correctly determined. The accuracy of determining the content body is 0.857, and the error is 0.143. Among the research test collection, 126 pages out of 146 labeled content pages are were correctly identified. The accuracy of determining the content body is 0.863, and the error is 0.137.

We investigated pages that have failed in the content body detection. The algorithm failed because the content body nodes detected by our algorithm were not exactly the labeled nodes. However, the detected nodes were still holding the content body. The example failure page in Figure 25 shows why we failed to determine the labeled content body. It is because it is difficult to judge the content body node for this example. The labeled outer rectangle border shows what is labeled and it includes the navigation parts in the right side. The inner rectangle border shows what the algorithm determined as a content body, and it excludes the title part of the article. However, there is no node that is holding both title and the article content in this example page.

From this analysis, we find that our algorithm performs accurately in most cases,

HOME PAGE MY TIMES TODAY'S PAPER VIDEO MOST POPULAR TIMES TOPICS

The New York Times
Tuesday, February 26, 2008

Multimedia/Photos


WORLD U.S. N.Y. / REGION BUSINESS TECHNOLOGY SCIENCE HEALTH SPORTS OPINION AI

Identity Theft is on the rise

Are you Protected?

Sarkozy Wins in France and Vows Break With Past

By ELAINE SCIOLINO



Ruth Fremson/The New York Times

French citizens celebrate the victory of Nicolas Sarkozy in the presidential election at the Place de la Concorde in Paris.

Nicolas Sarkozy earned a decisive victory on Sunday over Ségolène Royal, keeping the right in power for another five years. Turnout was 84 percent of France's registered voters.

Voices: The voters | Sarkozy

Slide Show: French Voters Pick Sarkozy | Times Topics

VIDEO: Tracing the Path of the Poisoned

A poisonous solvent sold by counterfeiters and mixed into drugs has figured in mass poisonings around the world that killed thousands.


Related Article

SLIDE SHOW: Clemens's Career Continues

At age 44, Roger Clemens will resume his pitching career by returning to the Yankees.

SLIDE SHOW: Smiles and Flowers

Takashi Murakami's show at Gagosian Gallery




Interactive Feature

The Victims

Read profiles of the men and women killed in the shootings at Virginia Tech and share your thoughts and memories.

Related Article



Interactive Feature

'Amazing Girls'

The achieving, ambitious and confident students of Newton North High School.

Related Article



Interactive Feature

INTERACTIVE FEATURE
Faces of the Dead

Fig. 24. One way to resolve this failure is to extract informative elements from the pages because the failed pages do contain the representative images and text, or we will be able to refine our algorithm and metrics to determine these pages as index pages. One possibility is that after the algorithm determines the content body, it ranks the content body node to determine whether it is the true content page or not by checking the link threshold.

but that sometimes it is hard to judge which node to identify as a content body that holds informative element in its sub-tree.

5. Experiment Results: Informative Image Detection

Our algorithm correctly detected 222 images out of 237 pages that are labeled from the news collection, and 128 images out of 146 pages. All of the pages from which the algorithm did not extract informative images are text-only article pages and no label for informative images as well. Therefore, we could determine all of the informative images to represent the labeled article pages.

The performance of the image detection algorithm is better than the content body detection algorithm because the error in the content body detection is not really misidentifying the content body, as explained in the previous section. The performance of informative image detection demonstrated that the representative images all reside in the sub-tree of the content body detected from the routine of the algorithm in the previous section. This means that from the correctly recognized content pages, we can extract images and contextual text accurately.

As both the informative images and text reside in the sub-tree of the informative content, in the same document context, we can associate the image with the text from the content body node to form the image+text representations.

D. Discussion

The performance of the algorithm is promising. We found over 90% recall and F1 using all of the news collection, research collection, and total test collection. However, the precision was on average 85% which is not as good as the recall and F1. It is because the algorithm fails to categorize on the index pages that contain substantial

BBC NEWS OPEN The News in 2 minutes News services Your news when you want it

News Front Page Last Updated: Wednesday, 10 January 2007, 02:09 GMT
E-mail this to a friend Printable version

Chavez accelerates on path to socialism

By Nathalie Malinarich
BBC News

Venezuelan President Hugo Chavez had always said that with his new term in office, beginning on 10 January, the socialist revolution would start in earnest. And, after his resounding victory on 3 December, he has wasted no time.

Before even being sworn in for the third time, Mr Chavez has said that he wants to merge all his coalition partners into a single party, warned he will not renew an opposition TV channel's licence and announced he will nationalise key businesses.

He has also called on the National Assembly to give him the power to rule by decree for a year and replaced his Vice-President, Jose Vicente Rangel, seen as a key figure in his previous administration.

While some of the announcements themselves have not come as a complete surprise, for many, the intensity and pace of the change has.

Surprises'

Exactly what the so-called deepening of the Bolivarian Revolution - named in honour of the 19th Century independence hero - would entail was not made clear during the presidential campaign.

Whatever its shape, the notion of the socialist days to come fills Mr Chavez's supporters with hope and his opponents with dread.

With each speech, Mr Chavez gives more details of what he plans to do.

Swearing in his cabinet two days before his own inauguration, Mr Chavez explained that the new era would be backed by "five engines", which would:

- allow him to rule by decree for 18 months
- lead to socialist constitutional reforms
- reinforce popular education
- change the "geometry of power" or the way political, social, economic and military power is distributed across

VENEZUELA UNDER CHAVEZ
KEY STORIES
 ▶ Chavez bid for more state control
 ▶ Chavez wins re-election
 ▶ Campaigns end in Venezuela poll
 ▶ Chavez holds final election rally
 ▶ Opposition rallies in Venezuela

FEATURES AND ANALYSIS
Full steam ahead
 President Chavez prepares to extend his socialist revolution as he starts a new term.

▶ Viewpoints on Chavez
 ▶ A nation divided
OPEN City of contrasts
 ▶ Photo journal: Barrio life
 ▶ Middle class feels the squeeze
 ▶ In pictures: Informal economy

BACKGROUND
 ▶ Hugo Chavez: Life and career
 ▶ Profile: Hugo Chavez
 ▶ Venezuela guide: Key facts
 ▶ Timeline: From Columbus on

HAVE YOUR SAY
 ▶ You asked President Chavez

TOP AMERICAS STORIES
 ▶ Deadly shooting at US university
 ▶ Chavez hosting summit on energy
 ▶ Eight killed as storms batter US

News feeds

MOST POPULAR STORIES NOW
MOST E-MAILED MOST READ

- 1 Deadly shooting at US university
- 2 In pictures: Virginia shootings
- 3 Gere kiss sparks India protests
- 4 Virginia shootings: Eyewitness accounts
- 5 Day in pictures

▶ Most popular now, in detail

Fig. 25. An example page that failed in determining the content body node. The outer rectangle border shows what is labeled as the content body block, and the inner rectangle border shows what the algorithm determined as the content body.

information about the hyperlinks. The algorithm struggles with these cases is not surprising. Thus, future research will work on further defining the semantics of these hybrid content-full index pages. Then, we can work on recognizing such pages, and their components, and then extracting the useful information from them.

There is a similar related work by Song *et al.* that determines the important block from web pages using the Support Vector Machine (SVM) and visual features which are the page rendering information in the browser [57]. The visual features are extracted by the VIPS algorithm, which is dependent on the Internet Explorer and Windows platform. Thus, their algorithm can only run on Windows platform, and as it needs to download and render the page, it adds more complexity to the algorithm. Our algorithm develops a decision tree to best utilize the document structure, and also eliminates the dependency on a training data by developing a new decision rule for this specific problem. We could not compare the algorithm performance with the same test collection because their test collection is not publicly available. Though, we could compare with the same measure with same type of collections, news pages. The performance of our algorithm with news collection was 95.3% F1 better than the work by Song *et al.*, 79% F1. This demonstrated that the visual features may helpful but not critical.

The human experience of search engines and other listings of contents will be enhanced by consistently representing entries using surrogates that combine images and text, because these make better use of human cognitive facilities than text-only representations. The present research presents a method for reliably extracting these preferred representations from web pages. In the next chapter, we move to enhance a new representation for collections, the composition space.

CHAPTER VII

RESULTDISTRIBUTOR: GENERATING DIVERSE INFORMATION IN
MIXED-INITIATIVE COMPOSITION

To support people better in experiencing information, enhancing the representation for collections is as important as the representation for individual documents, which is addressed in the previous chapter. We develop new structures to improve our new representation of collections, the composition space, which facilitates the human experience of visual connections among surrogates. To ensure diverse information from multiple sources to be presented evenly in the composition space, we developed a new control structure, ResultDistributor. ResultDistributor promotes the generation of diverse information from multiple sources in the mixed-initiative composition space.

Yahoo! Buzz is a web site that presents the top search queries made by users on the Yahoo! Search engine. This is calculated as a Buzz score; where each point is 0.001% of users searching Yahoo! on that day [65]. The top Buzz areas are modally categorized as Leaders (most searched subjects on that day) or Movers (greatest increase in Buzz score from one day to the next) on six topics: Actors, Movies, Music, Sports TV and Video Games. Thus, people who are interested in popular media can use Buzz to access the day's most popular topics. Considering the choices offered by the Buzz site as a whole, with many searches in each category, results in a large set of potential information resources for the human to browse through. With the Yahoo! Buzz interface, for each category, the participant can navigate from the list of top search queries to the top 15 search result pages and then to actual result documents, giving them a taste of the day's most popular content on the Internet.

For each Buzz topic and modality, combinFormation receives the set of the top 15 search queries. To represent this collection of searches to the participant, the

system will need to retrieve the first 15 result documents for each search. If the system naively processes searches, and then simply downloads result documents as the results arrive, the order of arrival of documents will be heavily biased to the first searches that are issued. Thus, we develop a refined control structure for generative searching and browsing. It is a new system structure, the ResultDistributor, which interleaves multiple searches to process large collections of searches and documents in the order that makes sense for the participant. With the integration of the new structure into combinFormation, a single composition space serves to represent and connect the 225 documents that result from each Buzz area. Further, multiple areas can be combined. Participants experienced combinFormation with this structure as easier to use and more interesting, entertaining, and were able to explore more diverse information than the typical browser interface. The research in this chapter is published in ACM Multimedia 2007 [66].

A. ResultDistributor

combinFormation can be used to issue multiple searches at one time, in order to combine and integrate results [23, 67, 24]. The Buzz scenario is an extreme case of search combination. We combine searches with separate queries to search engines, instead of combined “OR” searches, in order to meet the participant’s information needs. Typical search engines allow “OR” searches to gather different information in one space, but they show the results of the “OR” searches in page rank [68] order of the total set of search results, without giving equal priority to each search query the participant initiates. Example search results of a query, “apple OR orange”, in Figure 26 show that most of search results are from “apple” and few are from “orange”. Because “apple” is more well-known company than “orange” and many web pages are

referring the “apple” sites, it is obvious that “apple” sites have higher rank value than other “orange” sites. However, what participants expect when they explore multiple searches are distributed results from all queries.

What is the appropriate structure for processing multiple concurrent searches? In combination, concurrent threads retrieve search results. If we naively queue searches, and then simply download result documents as the results arrive, the order of arrival of documents will be heavily biased to the first searches that are issued. Another problem is that the connection and read times vary across result pages. Inasmuch as the participant wants to compare and integrate information from across the searches, this behavior is unacceptable. We need a control structure that gives information from each search equal weighting in the retrieval order and presentation. The idea is to balance the prioritization of bandwidth and CPU allocation across searches and result pages.

The ResultDistributor structure is developed to combine results from multiple searches equally. It processes multiple searches like a round-robin scheduling algorithm. It assigns equal priority to each search and handles search results in order. For example, the first result from the first search query will be processed equally with the first result of the second search query, even though the page rank order of the two search results are different. This control structure adjusts the order of downloading and processing in breadth-first manner to even out the searches’ media contributions.

The first step performed by the ResultDistributor is to request all of the searches, in order (see Figure 28). The ResultDistributor builds a data structure dynamically based on the number of search queries and the number of search results returned by the search engine(s) (see Figure 27). The number of search results may vary across different search queries. This will be recognized while processing searches, so the ResultDistributor can be aware of how processing proceeds for each search, and

Google [Web](#) [Images](#) [Video](#) [News](#) [Maps](#) [more »](#)

apple OR orange [Advanced Search](#)
[Preferences](#)

Web Results 1 - 10 of about 396,000,000 for **apple OR orange**. (0.08 seconds)

Apple
 Official site of **Apple, Inc.** [Stock quote for AAPL](#)
www.apple.com/ - 32k - Mar 28, 2007 - [Cached](#) - [Similar pages](#)
[iPod+iTunes](#) - www.apple.com/itunes/
[The Apple Store \(US\)](#) - store.apple.com/
[iPhone](#) - www.apple.com/iphone/
[Get a Mac](#) - www.apple.com/getamac/
[More results from www.apple.com »](#)

Apple - QuickTime
Apple's free media player supporting innumerable audio and video formats. The pro version includes an abundance of media authoring capabilities.
www.apple.com/quicktime/ - 7k - [Cached](#) - [Similar pages](#)

The Apple Store (U.S.)
 Manufacturer site; shop online for iBooks, PowerBooks, eMacs, iPods, and accessories. Also offers product support and downloads.
store.apple.com/ - 48k - Mar 28, 2007 - [Cached](#) - [Similar pages](#)

Orange UK Home Page
Orange UK offers search, news, entertainment and more. Get great mobile phone deals and broadband packages in the **Orange Shop**.
www.orange.co.uk/ - Mar 28, 2007 - [Similar pages](#) **result from orange**

Apple - Support
 Support for most **Apple** products provided by **Apple Computer**.
www.info.apple.com/ - [Similar pages](#)

Orange.com
 Supply a range of services and products for the mobile telephone market. Includes how to buy, links to individual country sites, investors information, ...
www.orange.com/ - 7k - [Cached](#) - [Similar pages](#) **result from orange**

Apple Developer Connection
 Provides news and technical information for **Apple Developers**.
developer.apple.com/ - [Similar pages](#)

Battery Exchange Program iBook G4 and PowerBook G4
Apple has determined that certain lithium-ion batteries containing cells manufactured by Sony Corporation of Japan pose a safety risk that may result in ...
support.apple.com/batteryprogram - [Similar pages](#)

Apple - Start
 Download iTunes 7.1 **Apple TV** syncing, full-screen Cover Flow, ... Walter S. Mossberg and Katherine Boehret review **Apple TV** for the Wall Street Journal, ...
livepage.apple.com/ - 37k - Mar 28, 2007 - [Cached](#) - [Similar pages](#)

iTunes and Windows Vista
Apple has released a new version of iTunes that addresses a number of ... **Apple** recommends Windows Vista customers who own an iPod install this software ...
docs.info.apple.com/article.html?artnum=305042 - 19k - Mar 28, 2007 - [Cached](#) - [Similar pages](#)

Sponsored Links

Apple Store™
 Buy **Apple** Hardware and Software at the Source. Many Orders Ship Free.
www.store.apple.com
 Texas

Apple Orange
 Looking for **Apple Orange**? Buy direct from sellers and save.
www.eBay.com

Go ooooooogoo g l e ▶

Result Page: [1](#) [2](#) [3](#) [4](#) [5](#) [6](#) [7](#) [8](#) [9](#) [10](#) [Next](#)

Fig. 26. Google's OR search results from a query, "apple OR orange". It only retrieves two results about "orange" among the first ten search results.

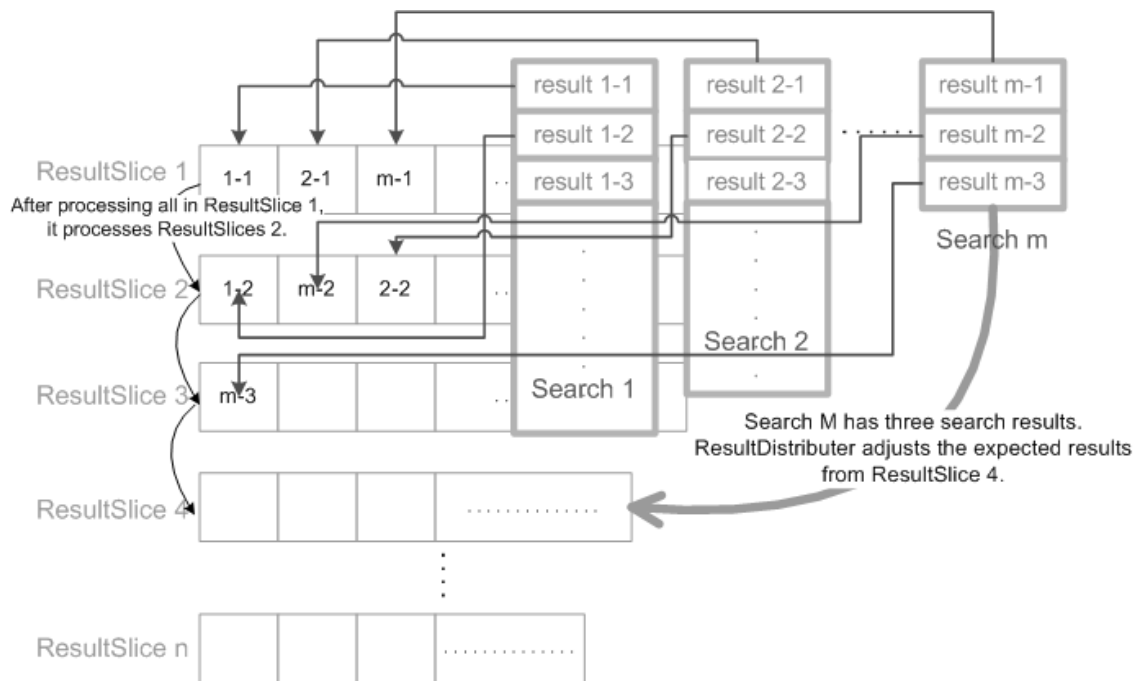


Fig. 27. Three concurrent processes in the ResultDistributor: (1) Process Search: add results into the appropriate ResultSlice; (2) Process Result: download and extract media from results in the ResultSlice. Move to the next ResultSlice after it finishes all results in the current slice; (3) Adjust the expected number of results in each ResultSlice.

Input : Search Queries (m)

Output : Distributed Media extracted from Results

Procedure : Start \rightarrow Process_Search(Q_i) where $0 \leq i < m$
 Start \rightarrow Process_Result(a, b) where $a=0, b=0$

```

Process_Search( Q ) {
    i=0;
    while( !downloadDone ) {
        Result = Q.parseNextResult();
        ResultSlice[i].add(Result);
        i++;
    }
    doneResult.add(i);
}

Process_Result( a, b ) {
    if( b < ExpectedNum(a) ) {
        Page = ResultSlice[a].get(b);
        Page.extractMedia();
    } else {
        a++; b=0;
        Process_Result(a, b);
    }
}

ExpectedNum( n ) {
    if( doneResult.remove(n) != null )
        return (-- m );
    return m;
}

```

Fig. 28. Pseudo code for the ResultDistributor.

when it needs to terminate. While it processes searches, results are inserted into a ResultSlice structure. The ResultSlice is a data structure that contains results of the same ordinality from multiple searches. The first results of all searches are inserted into ResultSlice 1. The second search results are inserted into ResultSlice 2, and so on. As all searches have been queued and they are being processed, processing of the first ResultSlice begins. Multiple concurrent worker threads are processing multiple searches; it is not deterministic which search results will be downloaded and processed first. Thus, results inserted into a single ResultSlice will be processed by the worker threads in a first-come, first-served manner. However, the next slice will not begin until processing the current slice completes.

The worker threads process search results starting from ResultSlice 1. They move progressively to the next ResultSlice as they finish processing the expected number of search results in the current ResultSlice. Let's assume that there are m searches, as in Figure 27. Consider a state in which the threads are processing ResultSlice 1, and it contains only 3 results. In this case, the expected number of search results in ResultSlice 1 is m . The current size of the ResultSlice is 3, so the ResultDistributor will wait until it finishes processing the remaining $m-3$ results in ResultSlice 1 before it goes to the next ResultSlice. The ResultDistributor sorts the order of downloading and processing of result set documents, concurrently with the download of the result sets themselves, as well as the documents.

It is important to continuously track the search processing status, because the order of result processing depends on the number of searches in each ResultSlice. When a particular search finishes processing all of its results, it notifies the ResultDistributor to adjust the expected size of appropriate ResultSlices (Figure 28). For example, 'Search m ' in Figure 27 has only three results, so the expected number of search results from ResultSlice 4 should be decreased by one. Even when a search

engine is unavailable, the rest of the searches can be processed under this structure. Either each read completes normally or exceptions, such as connection timeout, of problem search pages cause notification to the ResultDistributor.

The ResultDistributor structure supports multiple searches either using a single search engine, or with a combination of different search engines. combinFormation can apply this structure in processing multiple queries from diverse search engines selected by the participant, such as Google, Yahoo, Yahoo image, Flickr, and del.icio.us.

1. Yahoo Buzz with ResultDistributor

The ResultDistributor processes Yahoo Buzz in combinFormation. This allows participants to browse popular media from all the top search queries in an area at one time. combinFormation reads the Yahoo Buzz RSS feeds, which provide the fifteen top search queries for each topic and using either Leaders or Movers. It processes fifteen search results from fifteen top search queries in a ResultDistributor structure. Thus, the system immediately handles 225 web pages at a time and extracts media from them. Figure 29 presents a combinFormation composition space that resulted from the topic ‘TV Leaders’ in Yahoo Buzz. All 15 search results are presented equally in the space. A participant can directly browse top media without clicking links.

B. Balancing the Search Processing with Results

When the user launches combinFormation with a set of queries, it processes the request by sending those queries to the search engines. Then, it sends those search pages to a multithread structure, the DownloadMonitor, so that they can be retreated and parsed to generate result pages. The DownloadMonitor is a queue that contains the pages requested for downloading. As those search pages are processed one by one,



Fig. 29. Browsing TV Leaders in Yahoo Buzz in combinFormation. Large texts in rectangle boxes are labels added to identify the search queries of underlying clustered media elements.

result pages that are linked from the search pages get requested to be downloaded to the DownloadMonitor. Thus, the DownloadMonitor downloads and processes all the search pages first before it processes result pages.

This is not a problem if a user inputs few search queries. However, if a user inputs many queries like the Buzz (15 search queries), the DownloadMonitor will download and process fifteen search pages first, and then process result pages. If Internet downloading speed is fast, this is not a big issue, however when the Internet is slow, users will experience "combinFormation is doing nothing". This is because result pages will not be processed until the 15 search pages have finished downloading, and during this time no surrogates will be presented in combinFormaiton. The result pages contain the information elements and media to be extracted, and search pages have only links to those result pages.

In order to solve this issue, we implemented a waiting pipeline in the ResultDistributor to the DownloadMonitor. The DownloadMonitor counts the search pages that have been queued, and if they have been queued over the control limit, search pages will be seated in the waiting pipeline in ResultDistributor. The queue count will be reset when a result page is queued into the DownloadMonitor. This control flow is to queue search and result pages in a balanced manner to the DownloadMonitor, so that users can experience information and media without waiting even when the Internet speed is slow or when they put many search queries.

C. ResultDistributor Evaluation with Mixed-Initiative Composition

To assess the efficacy of the structurally enhanced combinFormation, we conducted a controlled experiment in which participants engaged in a browsing and authoring task using Yahoo Buzz. For each study condition, participants used either combinForma-

tion or a typical toolset to browse and collect popular media. For the typical toolset, Firefox was used for browsing and Microsoft Word for collecting interesting media. For combinFormation, participants could browse and collect popular media with the same mixed-initiative composition tool. In the study, participants were asked rating and essay questions about their experiences. We also logged the URLs that they browsed using each toolset.

1. Study Apparatus

The Study apparatus involves components for presenting the Buzz, for presenting tasks and study conditions to participants, and for logging the participants' browsing activities.

For the experiment, we divided the six Yahoo Buzz topics into two subsets. This was to eliminate carry-over effects that could result from conducting the same task with same media, but using different toolsets in separate conditions. One subset contains Sports, Music, and Video Games; the other subset contains Actors, Movies, and TV. We created a front page for each subset and toolset, in which participants can select interesting topics for them to browse during the study. The front page contains six choices of the three categories with both Leaders and Movers. By brushing a topic with mouse over, the fifteen top search queries are activated to appear on the right side of the topic. Thus, during the study, participants could identify interesting topics for them to browse.

When participants clicked one of topics on the front page, it opened either combinFormation or Firefox in a new window, to present the fifteen most popular searches. The front page reads the live Yahoo Buzz RSS Feeds; however there was a JavaScript security problem of reading and parsing XML from the other host inside the browser. So, we created a servlet that reads Yahoo Buzz RSS with the topic as a parameter and

outputs a live RSS string. The front page uses this servlet output to create dynamic HTML pages, and present participants the everyday updated popular media.

We developed the study apparatus using a framework for conducting user studies that is implemented with Javascript, Java Servlets, and XML. We defined the current experiment steps, counter-balanced orders, and resources in an XML file, and deployed the study to a server. The study Servlet rendered appropriate experimental conditions for each participant. It recorded times, conditions and other data. As each participant finished the study, data was stored in a repository in XML format.

There were two logging servers that recorded URLs participants browsed during the study. One was an HTTPPostServer, which processed HTTP Post messages sent from the browser. We used the Greasemonkey Firefox extension [69] to create a client resident in the browser to log every URL the user browses. This is more efficient and effective than using a proxy server. During a browsing task, the browser logging client accumulates entries for each user browse action. When the task finishes, it sends an HTTP Post message of the logs and a participant id to the logging server. The server saves the logging data in XML files. The other server was a combinFormation logging server, which records all the participants' interactions using the system. The combinFormation logging data is also saved in XML files.

2. Participants

Seven participants were involved in the study. They were college students, and their majors were three computer science, two general studies, one electrical engineering, and one economics. Six participants had never used combinFormation before, while one had. We asked about their browsing and Internet usage patterns. Five of them said they browse and search more than once a day and two said daily. We also asked how often they browsed the Internet with no specific tasks (casual browsing);

five of them said daily and two of them said a few times a week. The demographic information showed that the students search and browse casually on a daily basis.

3. Procedure

Participants were seated in front of a computer system with two monitors. At the beginning of the experiment, the experimenter introduced what the study was about for approximately 2-3 minutes. Whenever the participants were not clear about the experimental procedure, they could ask questions. The experimenter also explained about what is combinFormation, and demonstrated how to use it. Then, participants started the experiment.

The study phases consisted of pre-questionnaire, core-study, and post-questionnaire. The pre-questionnaire asked about participants' background. Next, in the core-study, they performed two browsing and authoring tasks. The task was:

“The page below shows 3 topics: Music, Sports and Video Games. The leaders are top queries in Yahoo. The movers represent queries that have gained the most in the past day. Hold your mouse over a topic to see what queries the leaders and movers contain. When you have decided to proceed in searching a category, click on the leaders or movers to start the search. If the current task is using combinFormation, the application will start. If this task requires that you use a browser with Microsoft Word, please open Microsoft Word by clicking on the icon on the Desktop. Collect any pictures or text that you find interesting. Reference appropriate information. Develop a collection that you could show to others to show your interests. Arrange the collection in combinFormation or in Microsoft Word in a way that expresses your interests clearly. If you are unhappy

with what your searches are giving you, you may replace or mix searches with combinFormation by clicking another category. If you are using a typical browser, clicking another category will open a new window with new search. If the current task is using combinFormation, use the interest expression mechanism to get more relevant results. Express positive interest in useful things, and negative interest in unhelpful things.”

Each task in the core-study took around 15-20 minutes. There were intermediate questions after each task. Before conducting the task with combinFormation, there was a brief practice session with combinFormation. In this session, participants used combinFormation for few minutes to browse news. The study concluded with a post-questionnaire, which asked about their experience using the two different toolsets. The whole procedure lasted approximately an hour.

4. Study Design

A 2x2 within-subjects design was employed. The independent variable was the browsing and authoring toolset. They were ‘mixed-initiative composition’ (combinFormation) and ‘typical’ (Firefox + Microsoft Word). The order of using the two different toolsets was counterbalanced. Each condition was associated with a different subset of the Buzz; the subset was randomly assigned to the toolset, and these assignments were also counterbalanced.

5. Results

We compared empirical logging data between two toolset conditions. For mixed-initiative composition, we counted the entire source URLs associated with media elements they browsed during the task. The results from logging data show that

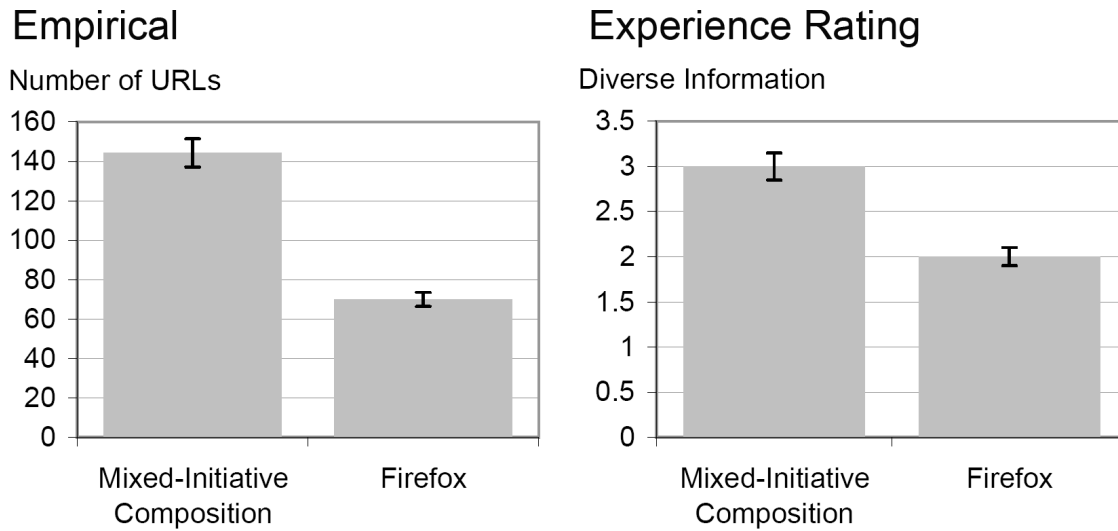


Fig. 30. Mixed-Initiative Composition enhances the diversity of media; Left: number of diverse URLs they browsed; Right: Participants' experience ratings about media diversity.

participants were able to browse media from more different web pages using combination than using the typical browser. On average, participants were able to browse 70 pages during the task with Firefox, and they browsed approximately 144 pages with mixed-initiative composition (see Figure 30 left). The difference between the numbers of browsing pages during the tasks was statistically significant [$F(1,6) = -6.071, p < 0.002$]. The result indicates that the ResultsDistributor-enhanced mixed-initiative composition lessens users' efforts and time of following links while experiencing popular media. It enhances the diversity of the media they experience.

In the intermediate questions, right after each browsing and authoring task, we asked each participant about how diverse was the information that they browsed. The answers were scaled from 1-5. We analyzed their answers, and the average rating score for mixed-initiative composition was 3.0, and that for the typical toolset was 2.0 (see Figure 30 right). With the Paired-Samples T-test, we found that participants sensed that they could browse more diverse information using mixed-initiative composition

(combinFormation) than using the typical browser [$F(1,6) = -3.240, p < 0.018$]. This user experience result corresponds with the empirical result from logging data, which confirmed the efficacy of the system for browsing diverse information.

In the post questionnaire, we asked participants about their subjective experience of the tools for easy to use, how much they liked it, finding interesting/relevant information, and entertaining (see Figure 31). Four participants answered that mixed-initiative composition was easier to use for browsing and collecting; one said the typical browser and Word, and two said both the same. The participant who selected the typical toolset explained the reason as their unfamiliarity using combinFormation. Participants who chose combinFormation expressed easiness of browsing different and diverse information without the extra work of following links.

P6: “With the combinFormation technique it allows you to look at topics in relationship to each other and it also gives you information from many different sites. That gives you the ability to see the information on one topic from different points and from sites you may never have found.”

P7: “It was sitting there right in front of you instead of you having to dig and dig for information. Any of the big time things in the world or with pop culture would pop up and it was everything you wanted to know.”

For the question about how much they liked the tools, three participants liked combinFormation better, two liked typical browser and Word, and the rest of two said both the same. Participants who enjoy browsing diverse popular media out in the world liked combinFormation. A couple of participants wanted to focus on specific information, instead of to engage in exploration; they like the control provided by the typical browser.

Then, participants were asked which tool they were able to find more interesting

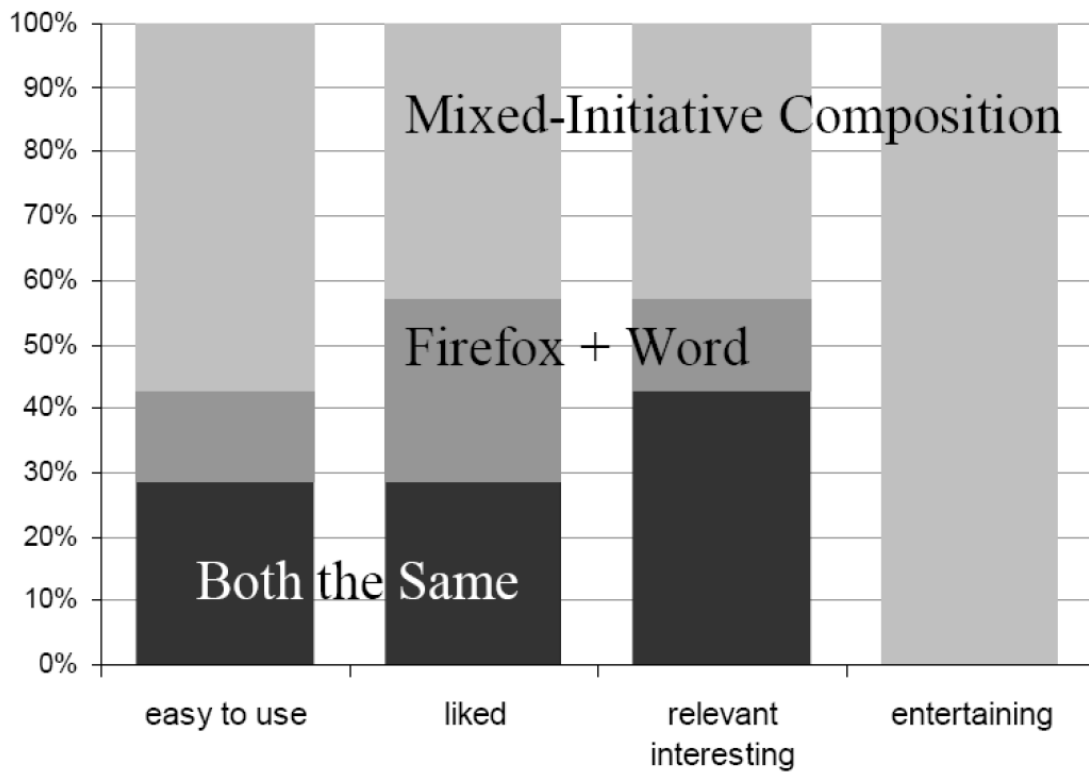


Fig. 31. Mixed-initiative composition is easier to use, more liked, finds more relevant and interesting media, and is more entertaining to experience than the typical toolset.

and relevant information with while browsing. Three participants answered combinFormation, one said the typical browser, and the rest three said both the same. Participants who found combinFormation useful said image media helped them browse and understand current popular issues. A participant who chose the typical browser said that using the browser he was more able to go directly to the relevant information he wanted.

P2: “I always think about these... When we browse popular media, I need to see just pictures not the texts... because it is easy to find something using that picture. In that point, combinFormation is easy to browse what I want.”

P1: “combinFormation is very useful to show current issues and just browse”

The last question was which tool was more entertaining. All participants selected combinFormation. They found combinFormation interesting and entertaining because they could look at many different things at the same time more visually.

P4: “combinFormation is more visual than the browser, it is very interesting”

P5: “It may help me to find current issues that many people talk about these days. I will surely use it when I want to find a ‘something new’.”

P6: “combinFormation was more entertaining because it gave you many things to look at the same time, giving you the ability to choose random topics that you might not think about.”

The results showed that combinFormation was more favorable and appealing to participants in the experience of browsing and authoring popular media than the typical

toolset. Participants enjoyed browsing visual representations of media extracted from different resources and authoring in a mixed-initiative composition space.

D. Background

What users need in experiencing multimedia is different from what they need for searching for finding information in general. In the area of multimedia searching, Jansen *et al.* noted that multimedia sessions and queries generally last longer than Web searches [70]. The longer queries indicate an increased cognitive load for multimedia searching. They found that multimedia requires greater interactivity between the user and search engine, in comparison with general Web searching. An increase in query and session lengths and in the number of viewed result pages in AltaVista Multimedia search logs indicates the greater interactivity. A term level analysis indicates the range of user information needs is broadening. Another study also reported a broadening of information needs [71]. Web users are searching for an increasing variety of multimedia topics.

Humans interact more and explore diverse information more in multimedia searching than in Web searching. The present research develops the ResultDistributor to broaden and diversify the media retrieved, and enhances interactive mechanisms of the composition space to improve the multimedia experience. Lew *et al.* find that we should focus as much as possible on the user who wants to explore media [72]. Decision makers need to explore an area to acquire valuable insight; thus, experiential systems which stress the exploration aspect are greatly needed.

E. Discussion

Using structurally enhanced combination, participants could experience diverse popular media and author creative media collections more easily and with a better sense of entertainment than with typical tools. Since diversity is a measure of creative ideation [1], this indicates that that combination's composition space promotes a more creative experience. Accessing diverse media is enabled by the ResultDistributor, which prioritizes assignment of network and CPU for processing of document links across search result sets. This ensures that the human participant receives a compositional representation that balances the contributions of searches and their results.

CHAPTER VIII

CONCLUSION

“What information consumes is rather obvious: it consumes that attention of its recipients. Hence a wealth of information creates a poverty of attention, and a need to allocate that attention efficiently among the overabundance of information sources that might consume it.”

As in the above quote by Herbert Simon, people have limited time and attention to acquire information from the web, which is a significant information resource. The importance of retrieving relevant and related information has resulted in the development of successful and powerful search engines, but the representation of information resources has not been improved accordingly.

By conducting a user study to develop understanding of collecting practices, we learned that people experience breakdowns. They are confronted with the problem of how to see, understand, think about, and remember significant elements within collections. Even though they built collections of elements that were useful, most information elements are not utilized in relevant contexts of need because of limits in human attention and memory. In order to address these problems, we develop two new representations for collections, the image+text surrogate and the composition space. The combination of image and text has been established to function as a cognitively better representation than a text-only representation by empirical research due to its visual characteristics that make use of complementary components of human working memory. Also, we demonstrated that using a composition space, people can browse more diverse information resources and develop more emergence of new ideas than using the typical linear list format [1]. Thus, the present research develops new methods to generate these representations that make better use of human attention.

Our first method is an information extraction algorithm that recognizes informative content within documents and extracts representative image and text clippings from the informative content to generate image+text surrogates. The second method is a new structure to process documents from multiple sources in a mixed-initiative composition space, ensuring that the generation of diverse information is balanced across different sources.

A. Image+Text Surrogate Extraction

We develop an information extraction algorithm to generate image+text surrogates from documents. The image+text surrogates provide a fine-grained summary of the document, so people can easily find and understand the document. Also, the visual cue of the image helps people quickly recognize and re-find the document. The image+text surrogates are formed by extracting a representative image from a given document and combining it with an informative textual passage that reveals what the associated part of the document is about.

We initially developed a surrogate classification algorithm to generate image+text surrogates. The initial algorithm generated surrogate candidates from all parts of a DOM tree, and classified informative or non-informative surrogate candidates with a supervised learning algorithm. The average classification performance was 80%. We developed ways to improve the algorithm. The new algorithm ranks sub-trees in the DOM to find the highest rank sub-tree that contains information; then it forms image+text surrogates only from that highest rank sub-tree. The new algorithm is a structurally more accurate and efficient way to address this problem. Thus, it produces better performance results: 90% accuracy on average. The new algorithm is developed as a decision tree, which utilizes the document's structure and eliminates

the dependency on training data by developing feature-based decision rules specific to this problem.

The new algorithm is structured in three stages. First, it determines whether the page is an index page or a content page. Then, the second stage recognizes the most informative sub-tree in each content page. The third stage extracts representative image and text surrogates from the sub-tree. We divide the algorithm into three stages in order to independently investigate the performance of each stage, enabling componentized improvements. In addition, we can enlarge the applicability of the algorithm to other information extraction and retrieval problems. For example, the output of the first stage can be applied and integrated to other page categorization algorithms and also search engines for crawling documents. Knowing the page type will be beneficial for search engines. If the search engines can recognize pages as index pages, they can determine their strategies to crawl the links in the pages for indexing more documents. Thus, search engines can utilize the first stage of our algorithm. Search engines can apply our algorithm to recognize the page type, determine the quality of hyperlinks by seeing the surrounding text near each link and the link importance itself, and finally decide to crawl more if necessary. The output of the second stage can be applied as a pre-processing stage for other algorithms to improve its performance. For example, a web page classification algorithm needs to find the similarity among documents e.g., [73]. The second stage of our algorithm can eliminate the noisy parts of documents that can disturb accurate similarity calculations. So, before applying similarity measures among documents, we can pre-process the document to determine the informative content in the document, and utilize only the informative content in further procedures. Further, the integration of all three stages can be utilized to represent the information resources in contexts such as search engines or digital libraries.

B. Implications for Search Engines

The problem addressed here is different from image search. Our problem is to visually represent the documents returned by everyday searches. In contrast, image search is designed to find the best matching images with queries that people form to specify what they are looking for. We are not building a new search engine, like image search. Rather, we are utilizing other document search engines by sending them queries and receiving search result documents. Then, we process those result documents to generate image+text surrogates to provide significant visual cues about the search results, so that people can easily find what they want. This can be the post-processing stage for the existing search engines to represent their result documents (see Figure 32). Or, while search engines are indexing documents, they can apply this algorithm to form surrogates, and index the surrogates along with the documents' other metadata. These surrogates can be one form of metadata, which can be utilized to speed up and enrich searching and browsing information.

While image search utilizes the text surrounding images and possibly image analysis techniques to find the relevant images to the queries, our research method recognizes the image that could represent the document best by determining the distance from the textual informative content in the DOM tree. We also utilize the text surrounding images to develop the semantics of images, but it is to determine the relevance of the images both to the document and to the query, not just to the query. Currently, we are using the text surrounding images only in the process of recognizing relevance to the informative portion of the document. For future work, it will be straight forward to measure the relevance of image surrogates to the query, so that when there are several informative surrogates in a document, we can select one based on the query relevance measure to best present what the user is looking for.

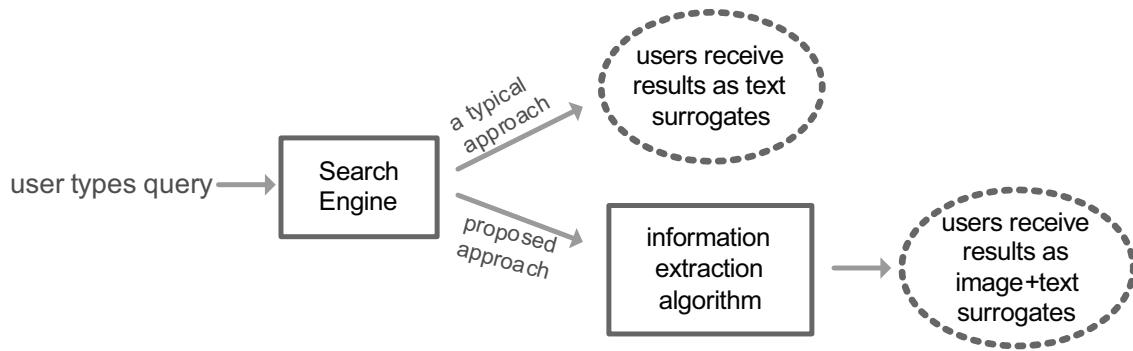


Fig. 32. A flow of how our information extraction algorithm can be used with a typical search engine to enhance the representation of results, and the overall user experience.

The image+text surrogates are formed from significant informative content of the document. If our algorithm cannot recognize the informative content from a document, it means that the document itself does not contain meaningful information. Search engines should not rank such documents highly, because the document will less directly fulfill people's information needs. It will consume their extra clicking efforts and time.


Even if our algorithm can determine the informative content, there will be the case that the algorithm cannot form quality surrogates with visual cues. This case arises when the document contains only text information or all the images in the document are not informative and representative. In this case, the rank of the document should perhaps slightly be reduced, because it means that the document does not contain informative content that is optimized for human cognition. (But the user may need these documents depending on what alternatives are available.) Thus, the quality of the available surrogates can be further incorporated into the document ranking metrics for search engines. In addition, this research can be supplemented by developing a method to generate alternative image representation for text-only informative documents, such as a screenshot thumbnail.

C. Further Improvements: Recognize Index-Content Pages

While we were analyzing the performance evaluation results, we identified that the classification accuracy of index pages was lower than that of content pages. Thus, we examined the index pages that were not correctly classified. We found that there are index pages that are somewhat like content pages, but they still function as index pages. These index pages contain substantial informative descriptions about other documents, along with hyperlinks to those documents. These are the borderline cases that our classification algorithm cannot recognize correctly, because the page types are recognized by the existence of informative content. We call these borderline pages index-content pages. A content page usually consists of an article about a single particular topic. An index-content page consists of a set of representations of other documents (see Figure 33). Each such representation in the index-content pages is usually formed with title text surrounded by a hyperlink, a thumbnail image, and a paragraph of description. The thumbnail image is associated with the same hyperlink. We recognize this representation as exactly providing the ingredients for forming image+text surrogates. Each index-content page consists of a set of these image+text surrogates.

As our initial idea of classifying pages into index or content develops a decision rule based on the existence of the substantial informative content, the index-content pages cannot be classified correctly with the current rule. We need to develop a second level rule to recognize these borderline cases.

Future work will address these borderline cases by developing a method to recognize the different patterns of the content body that arise between index-content pages and content pages. As the index-content pages contain a set of surrogates with a similar format, we can again apply a pattern recognition approach to determine




National Science Foundation
WHERE DISCOVERIES BEGIN

SEARCH

HOME | FUNDING | AWARDS | DISCOVERIES | **NEWS** | PUBLICATIONS | STATISTICS | ABOUT | FastLane


News




- [News](#)
- [News From the Field](#)
- [For the News Media](#)
- [Special Reports](#)
- [Research Overviews](#)
- [NSF-Wide Investments](#)
- [Speeches & Lectures](#)
- [NSF Current Newsletter](#)
- [Multimedia Gallery](#)
- [News Archive](#)

Overviews of NSF Research


Showing: 1 - 12 of 12



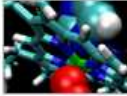
[Arctic & Antarctic](#)
 Earth's polar regions may seem like mirror images. But each is a unique environment with features we are still trying to understand, and each affects the rest of the globe.
 Date Updated: January 31, 2005




[Astronomy & Space](#)
 Astronomy may well be the oldest science of all, seeking answers to questions such as: "Where did it all come from?" and "Are we alone?"
 Date Updated: January 31, 2005




[Biology](#)
 A bat, a mushroom, a blade of grass--they're easy to identify as "life." But what about a cold virus or mold? Biologists are life's detectives, discovering what "alive" really means.
 Date Updated: January 31, 2005




[Chemistry & Materials](#)
 Most of what you touch, taste, hear or smell every day in the modern world is the direct or indirect result of research in chemistry or materials science - an effort that never ends.
 Date Updated: January 31, 2005




[Computing](#)
 The Internet, Google, and web browsers show how past progress in computing affects our daily lives. The cutting-edge systems under design now will have an enormous impact on society--and science.
 Date Updated: January 31, 2005




[Earth & Environment](#)
 Our planet gives up its secrets slowly. But every year we learn more about its oceans and air, its restless continents, its myriad ecosystems, and the way living things interact with their environments.
 Date Updated: January 31, 2005



[Education](#)
 Learning how people learn, while also supporting the very best ideas and students in U.S. science, engineering and mathematics are essential goals in today's changing world.
 Date Updated: January 31, 2005



[Engineering](#)
 Engineers bridge the gap between what the mind can imagine and what the laws of nature allow. They work at the outermost frontiers of electronics, manufacturing, bioengineering, structural design and more.
 Date Updated: January 31, 2005



[Mathematics](#)
 Mathematics is the natural language of science and engineering. It is about numbers, shapes, symmetry, chance, change and much more. Mathematics is deeply interwoven into all of modern life.
 Date Updated: January 31, 2005

Fig. 33. An example of an index-content page which could be recognized as a set of surrogates. Each surrogate consists of a title within a hyperlink, a thumbnail image, and a descriptive paragraph of text.

the surrogate format and the repetition of the format inside the content body of the index-content pages. We can also utilize the document structure by investigating how the link node is structured with informative content inside the content body sub-tree. Using the link threshold and patterns within the sub-tree of the content body node, we will be able to develop a new decision rule to differentiate the content body node of the index-content pages from that of the content pages.

After we can recognize such index-content pages, we can also develop methods to extract those surrogate entries. As such pages contain the surrogate entries with informative images and texts, we can form image+text surrogates for the hyperlink documents with less computational complexity. For example, if an index-content page contains six surrogate entries, we can form six image+text surrogates by processing the one index-content page to extract those surrogate entries. Otherwise, we need to process six documents to form six image+text surrogates. Extracting the surrogate entries from these index pages will be beneficial to interactive systems like combinFormation because the system can generate many informative surrogates by processing a single page. Similarly, this will help search engines to find many informative documents in addition to the surrogates for the document by processing a single page, which is more efficient and effective.

Even with these borderline cases, the performance of the algorithm is 90% accuracy on average and over 90% recall and F1, which proved the efficacy of the algorithm. Because of the misclassification of the index-content pages, the precision of the algorithm is lower than the recall and F1, 86% on average. The performance results demonstrated that the algorithm can already automatically generate image+text surrogates from most of documents accurately. Further, the results suggest that when the algorithm incorporates the solutions for the borderline cases, the precision of the algorithm will become higher, and will produce more accurate performance overall.

D. Increasing Diversity in Mixed-Initiative Composition Space

Then, we move to our second new representation for collections, the composition space. A composition space is an environment where people can compare, connect and relate information visually in a collage form using direct manipulation. In a mixed-initiative composition space, people can manipulate information based on their understanding while generative agents continuously provide visual information. This representation is designed to support information discovery, rather than simple finding of a single fact. To ensure generation of diverse information in the mixed-initiative composition space, we develop a new structure to control the processing order of documents from multiple sources to be balanced across, like a round-robin scheduling algorithm. User evaluation demonstrated that the participants were able to browse more diverse information with the ResultDistributor-enhanced composition space. Participants also found it easier and more entertaining to browse information in this modality.

By seeing the diverse image+text surrogates generated in the ResultDistributor-enhanced composition space, people can find, compare, and connect different information in collections without clicking links or switching different window views. This representation can facilitate the utilization of collections that may contain serendipitous or important information. Since diversity is a measure of creative ideation [74], this indicates that the ResultDistributor-enhanced composition space promotes a more creative experience. In this new representation, people can experience more diverse visual information, and manipulate the information according to their own way of thinking and understanding. Thus, it will promote information discovery [75]. The goal of information discovery tasks is not just finding information. It involves the emergence of new ideas in the context of the stimulus of found information. The

current search engine interfaces support people to find particular documents or information, but don't directly support people in developing ideas using found information. Thus, providing a new representation, the mixed-initiative composition space, additionally for search results will promote people to experience information discovery while finding information.

E. Summation

This research develops new visual representations for collections, an image+text surrogate and a composition space. The visual representation of documents, image+text surrogates can be applied in current search engine interfaces using the linear list format. The visual cues will help people to find information with better use of cognitive resources. This is a small step of transforming interfaces both for people and search engines, but we anticipate a big impact on people's experiences. The second representation, the composition space, advances further to help people to manipulate information and see connections among different and diverse information. This is a larger step of changing the paradigm for searching and collecting experiences. This interface will support people to think about and understand information in different ways, which can promote people to develop new ideas using found information. The composition space is especially beneficial for people engaged in information discovery tasks.

By applying these new visual representations to the interfaces of people's everyday means for interacting with information such as search engines, popular web sites, file directory explorer, and digital libraries, people will be able to see, understand, think about, and remember information more efficiently and effectively, with visual cognition. This will change people's everyday experience with information to be more

enjoyable and pleasant. Further, the new representations provide an interactive affordance to promote people to manipulate, think about, and develop new and different ideas among found information. This transcends the limits of current interfaces.

REFERENCES

- [1] A. Kerne, E. Koh, S. Smith, H. Choi, R. Graeber, and A. Webb, “Promoting emergence in information discovery by representing collections with composition,” in *Proc. ACM SIGCHI Conference on Creativity & Cognition*, 2007, pp. 117–126.
- [2] W3C, “Document Object Model (DOM) Level 2 Core Specification,” 2000. [Online]. Available: `\url{http://www.w3.org/TR/2000/REC-DOM-Level-2-Core-20001113/}`
- [3] T. Phelps and R. Wilensky, “Robust intra-document locations,” *Computer Networks*, vol. 33, no. 1-6, pp. 105–118, 2000.
- [4] A. Webb, *Statistical Pattern Recognition*, 2nd ed. New York: John Wiley & Sons Inc, 2002.
- [5] A. Kerne, E. Koh, B. Dworaczyk, J. Mistrot, H. Choi, S. Smith, R. Graeber, D. Caruso, A. Webb, R. Hill *et al.*, “combinFormation: A mixed-initiative system for representing collections as compositions of image and text surrogates,” in *Proc. 6th ACM/IEEE-CS Joint Conference on Digital Libraries*, 2006, pp. 11–20.
- [6] M. Burke, *Organization of Multimedia Resources: Principles and Practice of Information Retrieval*. Hampshire, UK: Gower, 1999.
- [7] Miniwatts Marketing Group, “Internet usage statistics - the big picture,” <http://www.internetworldstats.com/stats.htm>, last visited 11/20/2006.

- [8] L. Rainie, "Internet: The mainstreaming of online life," *Pew Internet and American Life Project*, 2005. [Online]. Available: [\url{http://www.pewinternet.org/pdfs/Internet_Status_2005.pdf}](http://www.pewinternet.org/pdfs/Internet_Status_2005.pdf)
- [9] D. Sullivan and I. M. Group, *Search Engine Watch*. INT Media Group, last visited 11/20/06, <http://searchenginewatch.com/showPage.html?page=2156461>.
- [10] R. White, I. Ruthven, and J. Jose, "Finding relevant documents using top ranking sentences: An evaluation of two alternative schemes," in *Proc. ACM SIGIR Conference on Research and Development in Information Retrieval*, 2002, pp. 57–64.
- [11] S. Dumais, E. Cutrell, and H. Chen, "Optimizing search by showing results in context," in *Proc. SIGCHI Conference on Human Factors in Computing Systems*, 2001, pp. 277–284.
- [12] T. Paek, S. Dumais, and R. Logan, "WaveLens: A new view onto internet search results," in *Pro. SIGCHI conference on Human factors in computing systems*, 2004, pp. 727–734.
- [13] W. Ding, G. Marchionini, and D. Soergel, "Multimodal surrogates for video browsing," in *Proc. 4th ACM Conference on Digital Libraries*, 1999, pp. 85–93.
- [14] B. Wildemuth, G. Marchionini, M. Yang, G. Geisler, T. Wilkens, A. Hughes, and R. Gruss, "How fast is too fast? Evaluating fast forward surrogates for digital video," in *Proc. ACM/IEEE-CS Joint Conference on Digital Libraries*, 2003, pp. 221–230.
- [15] A. Woodruff, A. Faulring, R. Rosenholtz, J. Morrisson, and P. Pirolli, "Using thumbnails to search the web," in *Proc. SIGCHI Conference on Human Factors*

- in Computing Systems*, 2001, pp. 198–205.
- [16] A. Baddeley, “Is working memory working?” *The Quarterly Journal of Experimental Psychology Section A*, vol. 44, no. 1, pp. 1–31, 1992.
- [17] A. Glenberg and W. Langston, “Comprehension of illustrated text: Pictures help to build mental models,” *Journal of Memory and Language*, vol. 31, no. 2, pp. 129–151, 1992.
- [18] A. Glenberg, “The indexical hypothesis: Meaning from language, world, and image,” *Working with Words and Images: New Steps in and Old Dance*, pp. 27–42, 2002.
- [19] R. Carney and J. Levin, “Pictorial illustrations still improve students’ learning from text,” *Educational Psychology Review*, vol. 14, no. 1, pp. 5–26, 2002.
- [20] R. Mayer and R. Moreno, “A Split-Attention Effect in Multimedia Learning: Evidence for Dual Processing Systems in Working Memory.” *Journal of Educational Psychology*, vol. 90, no. 2, pp. 312–20, 1998.
- [21] C. H. Smith, *The Oxford English Dictionary, Second Edition on CD-ROM, Version 3.1*, 2005.
- [22] A. Kerne, “Collagemachine: A model of "interface ecology",” Ph.D. dissertation, New York University, 2001.
- [23] Interface Ecology Lab, “combinFormation,” 2008, <http://ecologylab.cs.tamu.edu/combinFormation/>.
- [24] A. Kerne, E. Koh, S. Smith, A. Webb, and B. Dworaczyk, “Facilitating information discovery by collecting with mixed-initiative composition of image and text surrogates,” *ACM Transactions on Information Systems*, (in press 2008).

- [25] H. Simon, "Designing organizations for an information-rich world," *Computers, Communications, and the Public Interest*, pp. 38–52, 1971.
- [26] T. Winograd and F. Flores, *Understanding Computers and Cognition*. Norwood, NJ: Ablex Publishing Corp., 1986.
- [27] S. Jones, "The internet goes to college," *Pew Internet & American Life*, vol. 15, 2002. [Online]. Available: [\url{http://www.pewinternet.org/pdfs/PIP_College_Report.pdf}](http://www.pewinternet.org/pdfs/PIP_College_Report.pdf)
- [28] E. Koh and A. Kerne, "'I keep collecting': College students build and utilize collections in spite of breakdowns," in *Proc. European Conference on Digital Libraries*, 2006, pp. 304–314.
- [29] O. Bälter, "Strategies for organizing email," in *Proc. of Human Computer Interaction on People and Computers XII*, 1997, pp. 21–38.
- [30] D. Abrams, R. Baecker, and M. Chignell, "Information archiving with bookmarks: Personal web space construction and organization," in *Pro. SIGCHI Conference on Human Factors in Computing Systems*, May, 1998, pp. 41–48.
- [31] W. Jones, S. Dumais, and H. Bruce, "Once found, what then? a study of "keeping" behaviors in the personal use of web information," *Proc. the American Society for Information Science and Technology*, vol. 39, no. 1, pp. 391–402, 2002.
- [32] R. Boardman and M. Sasse, "'Stuff goes into the computer and doesn't come out': A cross-tool study of personal information management," in *Proc. SIGCHI Conference on Human Factors in Computing Systems*, April, 2004, pp. 583–590.
- [33] T. Malone, "How do people organize their desks?: Implications for the design of office information systems," *ACM Transactions on Information Systems (TOIS)*,

- vol. 1, no. 1, pp. 99–112, 1983.
- [34] S. Whittaker and C. Sidner, “Email overload: Exploring personal information management of email,” in *Proc. SIGCHI Conference on Human Factors in Computing Systems*, April 13–18 1996, pp. 276–283.
- [35] D. Barreau and B. Nardi, “Finding and reminding: File organization from the desktop,” *ACM SIGCHI Bulletin*, vol. 27, no. 3, pp. 39–43, 1995.
- [36] L. Suchman, *Plans and Situated Actions: The Problem of Human-Machine Communication*. New York: Cambridge University Press, 1987.
- [37] F. Shipman and C. Marshall, “Formality considered harmful: Experiences, emerging themes, and directions on the use of formal representations in interactive systems,” *Computer Supported Cooperative Work (CSCW)*, vol. 8, no. 4, pp. 333–352, 1999.
- [38] M. Czerwinski, D. Gage, J. Gemmell, C. Marshall, M. Pérez-Quñones, M. Skeels, and T. Catarci, “Digital memories in an era of ubiquitous computing and abundant storage,” *Communications of the ACM*, vol. 49, no. 1, pp. 44–50, 2006.
- [39] C. Marshall, S. Bly, and F. Brun-Cottan, “The long term fate of our personal digital belongings: Toward a service model for personal archives,” in *Proc. IS&T Archiving*, 2006, pp. 25–30.
- [40] E. Koh, D. Caruso, A. Kerne, and R. Gutierrez-Osuna, “Elimination of junk document surrogate candidates through pattern recognition,” in *Proc. ACM Symposium on Document Engineering*, 2007, pp. 187–195.
- [41] A. Arasu and H. Garcia-Molina, “Extracting structured data from web pages,” in *Proc. ACM SIGMOD International Conference on Management of Data*, June,

- 2003, pp. 337–348.
- [42] C. Chang and S. Lui, “IEPAD: Information extraction based on pattern discovery,” in *Proc. 10th International Conference on World Wide Web*, 2001, pp. 681–688.
- [43] S. Lin and J. Ho, “Discovering informative content blocks from web documents,” in *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2002, pp. 588–593.
- [44] W3C, “Cascading style sheets (CSS),” last visited 05/19/2008. [Online]. Available: `\url{http://www.w3.org/Style/CSS/}`
- [45] N. Rowe, J. Coffman, Y. Degirmenci, S. Hall, S. Lee, and C. Williams, “Automatic removal of advertising from web-page display,” in *Proc. ACM/IEEE-CS Joint Conference on Digital Libraries*, 2002, pp. 406–406.
- [46] C. Bishop, *Neural Networks for Pattern Recognition*. New York: Oxford University Press, 1995.
- [47] G. Crane and J. Rydberg-Cox, “New technology and new roles: The need for “corpus editors”,” in *Proc. 5th ACM Conference on Digital Libraries*, 2000, pp. 252–253.
- [48] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*. Harlow, England: Addison-Wesley, 1999.
- [49] National Institute of Standards and Technology, “Text REtrieval Conference (TREC),” <http://trec.nist.gov>, last visited 01/08/2008.

- [50] L. Slaughter, G. Marchionini, and G. Geisler, "Open video: A framework for a test collection," *Journal of Network and Computer Applications*, vol. 23, no. 3, pp. 219–245, 2000.
- [51] Z. O. Toups, A. Kerne, and A. Webb, "Composing service components with lexical scoping for little semantic webs: An expressive framework," submitted to Intl Semantic Web Conference 2008.
- [52] National Institute of Standards and Technology, "TREC Video Retrieval Evaluation (TRECVID)," <http://www-nlpir.nist.gov/projects/trecvid/>, last visited 01/29/2008.
- [53] INEX, "Initiative for the evaluation of XML retrieval," <http://www.is.informatik.uni-duisburg.de/projects/inex/index.html>, last visited 03/20/2008.
- [54] W. Dakka and L. Gravano, "Efficient summarization-aware search for online news articles," in *Proc. 2007 Conference on Digital Libraries*, 2007, pp. 63–72.
- [55] PSR Computer Consulting, "Newsblaster," last visited 01/29/2008. [Online]. Available: `\url{http://www.newsblaster.com/}`
- [56] Y. Liu, K. Bai, P. Mitra, and C. Giles, "TableSeer: Automatic table metadata extraction and searching in digital libraries," in *Proc. 2007 Conference on Digital Libraries*, 2007, pp. 91–100.
- [57] R. Song, H. Liu, J. Wen, and W. Ma, "Learning important models for web page blocks based on layout and content analysis," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 2, pp. 14–23, 2004.
- [58] D. Cai, S. Yu, J. Wen, and W. Ma, "VIPS: A vision based page segmentation algorithm," Beijing: Microsoft Research Asia, Tech. Rep. MSR-TR-2003-79, Nov.,

2003.

- [59] Mozilla, “DOM inspector,” <http://www.mozilla.org/projects/inspector/>, last visited 12/13/2007.
- [60] A. Quick, “JTidy,” <http://jtidy.sourceforge.net/>, last visited 01/08/2008.
- [61] C. Chang, M. Kayed, M. Girgis, and K. Shaalan, “A survey of web information extraction systems,” *Knowledge and Data Engineering, IEEE Transactions on*, vol. 18, no. 10, pp. 1411–1428, 2006.
- [62] A. Laender, B. Ribeiro-Neto, A. da Silva, and J. Teixeira, “A brief survey of web data extraction tools,” *ACM SIGMOD Record*, vol. 31, no. 2, pp. 84–93, 2002.
- [63] J. Chen, B. Zhou, J. Shi, H. Zhang, and Q. Fengwu, “Function-based object model towards website adaptation,” in *Proc. 10th International Conference on World Wide Web*, 2001, pp. 587–596.
- [64] A. Jaimes, D. Gatica-Perez, N. Sebe, and T. Huang, “Human-centered computing: Toward a human revolution,” *IEEE Computer*, vol. 40, no. 5, pp. 30–34, 2007.
- [65] Yahoo!, “Buzz: Frequently asked questions,” last visited 03/30/2007. [Online]. Available: `\url{http://buzz.yahoo.com/faq/}`
- [66] E. Koh, A. Kerne, A. Webb, S. Damaraju, and D. Sturdivant, “Generating views of the buzz: Browsing popular media and authoring using mixed-initiative composition,” in *Proc. ACM Conference on Multimedia*, 2007, pp. 228–237.
- [67] E. Koh, A. Kerne, and R. Hill, “Creativity support: Information discovery and exploratory search,” in *Proc. ACM SIGIR Conference on Research and Development in Information Retrieval*, 2007, pp. 895–896.

- [68] S. Brin and L. Page, “The anatomy of a large-scale hypertextual web search engine,” *Computer Networks and ISDN Systems*, vol. 30, no. 1-7, pp. 107–117, 1998.
- [69] A. Boodman, “Greasemonkey,” Firefox browser extension, <http://www.greasepot.net/>, last visited 04/13/2007.
- [70] B. Jansen, A. Spink, and J. Pedersen, “An analysis of multimedia searching on AltaVista,” in *Proc. 5th ACM SIGMM International Workshop on Multimedia Information Retrieval*, 2003, pp. 186–192.
- [71] A. Spink, B. Jansen, D. Wolfram, and T. Saracevic, “From e-sex to e-commerce: Web search changes,” *Computer*, vol. 35, no. 3, pp. 107–109, 2002.
- [72] M. Lew, N. Sebe, C. Djeraba, and R. Jain, “Content-based multimedia information retrieval: State of the art and challenges,” *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)*, vol. 2, no. 1, pp. 1–19, 2006.
- [73] D. Shen, Z. Chen, Q. Yang, H. Zeng, B. Zhang, Y. Lu, and W. Ma, “Web-page classification through summarization,” in *Proc. ACM SIGIR Conference on Research and Development in Information Retrieval*, 2004, pp. 242–249.
- [74] R. Finke, T. Ward, and S. Smith, *Creative Cognition: Theory, Research, and Applications*. Cambridge, MA: MIT Press, 1992.
- [75] A. Kerne and S. Smith, “The information discovery framework,” in *Proc. Designing Interactive Systems (DIS): Processes, Practices, Methods, and Techniques*, 2004, pp. 357–360.

VITA

Eunyee Koh received a B.S. in Computer Science and Engineering from Seoul National University in 2002, and worked for Motorola during 2002-2003. She received her Doctor of Philosophy in Computer Science from Texas A&M University in August 2008. Her research area includes human computer interaction and information retrieval, focusing on designing, architecting, implementing, and evaluating an information system, combinFormation. The research results in publishing papers in Joint Conference on Digital Libraries (JCDL), European Conference on Digital Libraries (ECDL), ACM Multimedia, ACM Computer Human Interaction, SIGIR, and ACM Document Engineering. She was a reviewer for IEEE Transactions on Human-Centered Computing and ACM Computer Human Interaction. In addition, she is actively involved in volunteering and mentoring work in the women in computer science group, and also she was a student volunteer in various conferences such as the JCDL, ACM CHI, and Grace Hopper Conferences.

Eunyee will join Adobe Systems Incorporated at San Jose starting June 2008 as a Computer Scientist. Her email is eunyee.koh@gmail.com, and her contact address is 301 Harvey R. Bright Bldg. College Station, TX 77843-3112.