Elimination of Junk Document Surrogate Candidates through Pattern Recognition

Eunyee Koh, Daniel Caruso, Andruid Kerne, Ricardo Gutierrez-Osuna
Interface Ecology Lab
Center for Study of Digital Libraries | Computer Science Department
Texas A&M University, College Station, TX 77843, USA
{eunyee, dcaruso, andruid, rgutier}@cs.tamu.edu

ABSTRACT
A surrogate is an object that stands for a document and enables navigation to that document. Hypermedia is often represented with textual surrogates, even though studies have shown that image and text surrogates facilitate the formation of mental models and overall understanding. Surrogates may be formed by breaking a document down into a set of smaller elements, each of which is 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 paper develops a pattern recognition based approach for eliminating junk while building the set of 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. We end by describing a new approach for hypermedia and the semantic web, which uses the DSM to define value-added surrogates for a document.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Selection process H.5.4 [**Information Interfaces and Presentation**]: Hypertext/Hypermedia – Navigation.

General Terms
Algorithms, Performance, Design, Human Factors

Keywords
surrogate, document surrogate model, navigation, mixed-initiatives, pattern recognition, quadratic classifier, principal components analysis, semi-autonomous agents

DocEng'07, August 28-31, 2007, Winnipeg, Manitoba, Canada.

1. INTRODUCTION
Representing large collections of documents to users in ways that facilitate understanding the essential meanings that the documents convey is a hard problem. This is a form of Vanevar Bush's problem which frames our field: there is too much information [4]. Surrogates are information elements selected from a specific document, which can be used in place of the original document [3, 25]. Most responses to search queries are represented in the form of lists of textual surrogates [14, 32, 35]. Yet, studies have shown that users prefer image and text surrogates and understand them more readily [10, 20]. Further, image and text representations facilitate the formation of mental models [13]. Building good image and text surrogates for a document is not simple and straightforward. One approach to this problem is to explicitly include image and text surrogates among the metadata that is specified for each document, just as abstracts are kept as textual representations. Image and text surrogates function as "boosters" [28] that add value to the process of content aggregation by promoting collection understanding [6].

Alternatively one may extract surrogates from documents through procedural methods. The nature of this task differs depending on the document format. Some digital libraries and semantic web repositories include a large number of HTML documents and sites [26, 27]. Extracting good visual surrogates from documents in this type of collection is complicated by the presence of *junk*, such as site navigational elements, which may not represent the document's meaning.

In addition to how individual surrogates are represented, another issue is how to represent collections. One approach to representing collections would be to use lists of image and text surrogates instead of pure text in the result sets that search engines return. An alternative approach is taken by combinFormation [18], a tool that facilitates the construction of surrogates and their spatial and visual composition in a mixed-initiative system [23]. Compositions are produced by a generative agent whose actions can be overridden and directed by the user. combinFormation uses surrogates in a variety of ways, such as changing the interest model used by the agent, visually combining surrogates to illustrate an idea or concept, and navigating to the original document the surrogate was selected from.

Imagine a space where interesting pieces of the most up-to-date information on your favorite topics are continuously discovered and presented to you. Now imagine if this space was full of advertisements, e-mail addresses, copyright notices, website navigation bars, etc. Sorting through and uncovering the information you are actually interested in becomes a difficult and

Figure 1. These pools of surrogate candidates have been manually separated into junk and non-junk to illustrate what we mean by junk, and how much of it there is.

cognitively expensive process. Unfortunately, in some simple exercises conducted using combinFormation, this is exactly what happened. In Figure 1, we have separated out all of these garbage elements. Having to perform this task every few seconds, as new information elements are presented, is a distracting task for the user. It would be better for the user if an application removed these elements automatically, freeing up her/his cognitive abilities for more important tasks such as processing the real information s/he is interested in. Automatically choosing the most informative candidates is a difficult task. combinFormation's prior surrogate candidate filtering was based on heuristic approaches. As we have seen, this approach performs poorly when we are operating on many HTML documents, since large portions of them contain non-informative elements such as advertisements, navigation menus, copyrights, etc.

In an ideal world, selected information elements give users an impression about the underlying meaning of the documents they come from. They also enable navigation to the complete document they represent. When chosen effectively, these elements function as representative surrogates for their containing documents. Rather than require a human being to perform this task, we need to discover how to design an agent that can perform it effectively.

Although the definition of junk is subjective, an ideal surrogate represents the ideas within a document. In the human experience of reading, navigation elements, advertisements, copyright notices, and mailing addresses do not effectively perform this function.

In this paper, we apply statistical pattern recognition techniques to a set of human judgments which have been systematized in the form of training data, in order to determine if any subsequently encountered surrogate should be discarded as junk. Our overall goal is to improve surrogate selection by increasing the number of junk surrogates that are correctly discarded.

2. RELATED WORK

Since users often see surrogates before the documents that they represent, they are used to make rapid decisions about whether to examine an information object in greater detail or not [15]. Marchionini *et al*. investigated the use of multimodal surrogates for video browsing [10, 31]. Their experiments compared users' performance and experience using different kinds of surrogates for digital videos. They engaged in a qualitative investigation of users' cognitive processes. Our research is also based on users' experiences of surrogates. We utilize human input in generating training sets, which ultimately drive pattern recognition algorithms. To account for diversity in human perspective, we incorporated surrogate junk identification judgments made by several researchers.

Prior research has developed valuable methods for modeling web page documents, defining useful features sets, and using the features to recognize structures within these documents. EXALG [1] 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 [1]. 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 [5] 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 [5]. 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 [24] 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) [33] as well as tables in order to structure formatting.

Rowe *et al.* [29] 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 [1, 5], 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 [29], 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 [1, 5]. 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 application of pattern recognition techniques. It is also an example of what we mean by a semi-autonomous process. The next sections describe the features we are using for identifying surrogates, the procedure for gathering training data, the algorithms for classification, and the results produced.

3. 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. [29], and EXALG [1], 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 1.

Table 1. Document Surrogate Features

Type	Features
Images	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* [29]. 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" [7]. 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 [29].

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.

4. DOCUMENT SURROGATE MODEL

The Document Object Model (DOM) is a tree-structured representation of a document [34] 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 [34]. 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, combinFormation 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

Table 2. 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

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.

5. 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 1, 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 [11].

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.

5.1 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 $\sum_i$ (i={*junk, non_junk*}):

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

where x is the feature vector associated with a given surrogate. Following the Maximum A Posteriori principle [11], surrogate x is classified according to the decision rule:

$$x \in \begin{cases} junk & if \;\; P(junk|x) > P(non\_junk|x) \\ non\_junk & otherwise \end{cases} \quad (2)$$

where *P(junk|x)* and *P(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 (2) can be expressed as:

$$x \in \begin{cases} junk & if \;\; \frac{P(x|junk)}{P(x|non\_junk)} > \frac{P(non\_junk)}{P(junk)} \\ non\_junk & otherwise \end{cases} \quad (3)$$

where *P(junk)* is the frequency of junk surrogates, also known as the *prior* probability, and *P(x|junk)* is the density of examples for the junk class, which by equation (1) we assume to be normally distributed. *P(non_junk)* and *P(x|non_junk)* are defined similarly. Merging equations (1) and (3) and taking natural logarithms:

$$x \in \begin{cases} junk & if \;\; g_{junk}(x) > g_{non\_junk}(x) \\ non\_junk & otherwise \end{cases} \quad (4)$$

where:

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i) - \frac{1}{2}\log|\Sigma_i| + \log P(\omega_i) \quad (5)$$

To build a quadratic classifier one need only compute the mean μ_i and covariance matrices $\sum_i$ 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) and (5).

5.2 CROSS-VALIDATION METHOD

The performance of the quadratic classifier is estimated by means of k-fold cross-validation [11,30]. 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) and (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.

6. EXPERIMENTS

6.1 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" surrogates 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] was seeded [22] 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*.

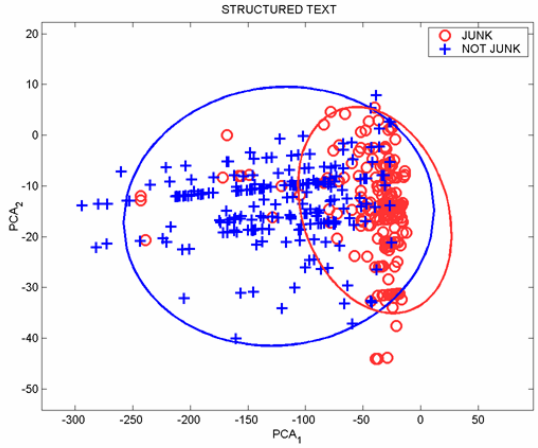

Figure 2. 2D PCA scatter plot of text surrogate candidates in the Structured Collection.

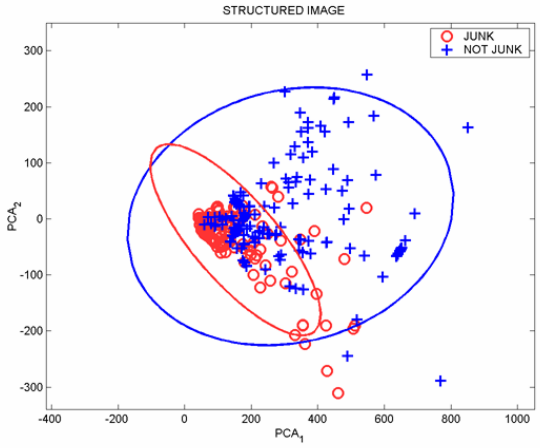

Figure 3. 2D PCA scatter plot of image surrogate candidates in the Structured Collection.

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 [21], 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 [22]. 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" [8]. 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.

6.2 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 2-8), 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.

6.2.1 THE STRUCTURED COLLECTION

Shown in Figure 2, 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 image surrogates, the distribution looks less Gaussian, and so the assumption of the classifier might be questioned. 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

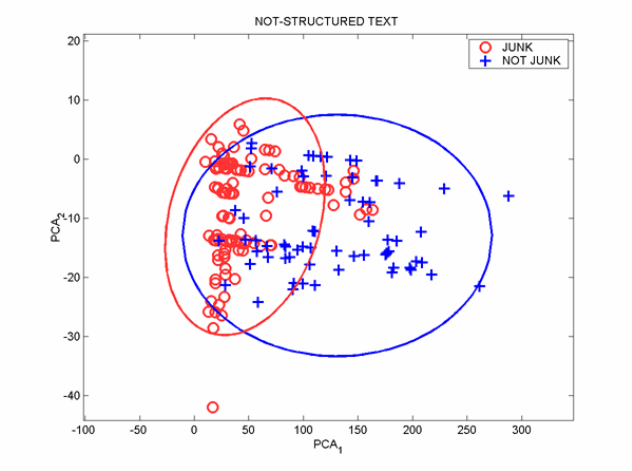

Figure 4. 2D PCA scatter plot of text surrogate candidates in the Non-Structured Collection.

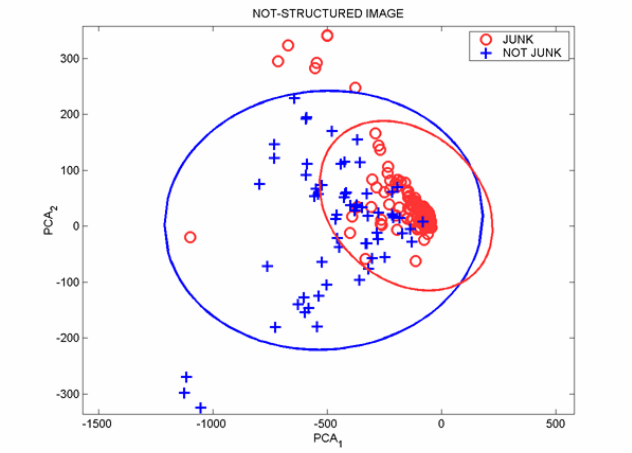

Figure 5. 2D PCA scatter plot of image surrogate candidates in the Non-Structured Collection.

than the text surrogate space. This result suggests that better performance may be obtained with alternative classification algorithms.

6.2.2 THE NON-STRUCTURED COLLECTION

The PCA projections for the Non-Structured Collection (Figures 4 - 5) 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.

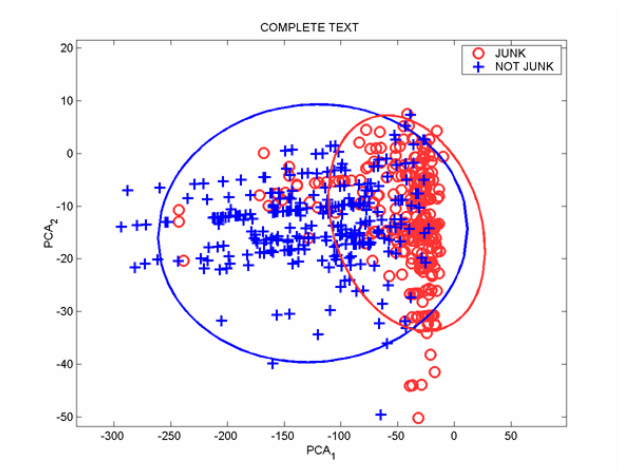

Figure 6. 2D PCA scatter plot of text surrogate candidates in the Complete Set

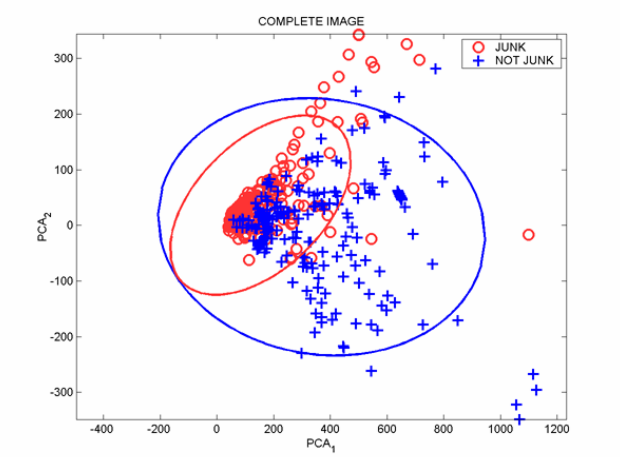

Figure 7. 2D PCA scatter plot of image surrogate candidates in the Complete Set

6.2.3 THE COMPLETE SET

The PCA projections for the Complete Set (Figures 6 - 7), 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 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.

7. DISCUSSION

7.1 PATTERN RECOGNITION

Performance measures for the three datasets and feature spaces are summarized in Table 2. The average performance of the quadratic classifier (80%) is significantly above chance level (50%), indicating that the selected features in Table 1 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.

The performance of the quadratic classifier may be improved by means of a robust estimate of mean and variance, so as to avoid sensitivity to outliers. Feature subset selection techniques [11, 30] could also be employed to determine a small subset of highly informative features from an initially large pool of candidate features. Improved performance may be obtained with a less restrictive classifier model. Such may be the case of semi-parametric approaches, such as Gaussian Mixture Models [16], or non-linear models such as Multilayer Perceptrons or Support Vector Machines [2].

7.2 DIGITAL COLLECTIONS

Beyond our classification of junk surrogate candidates, the development of the Document Surrogate Model structure is significant in itself. While this structure already plays a critical role in surrogate candidate creation and selection, use of the DSM for construction of surrogate nodes is only one possible application.

Like other “booster value-added surrogates” [28], the DSM can function as a type of single source metadocument [12]. Surrogates and their relationships could be authored rather than generated. Authored DSMs could potentially be a very compact structure with which to provide a basis for presenting collections of documents in digital libraries. DSMs could be authored in a manner similar to abstracts, which authors currently write so they can function as textual components of surrogates for research papers. Further work can investigate the potential role of such metadocuments in digital libraries. Formalized semantic XML structures for the DSM can be created, enabling these structures to be easily published, and utilized by collection representation systems.

In his famous work about the Memex, Vanevar Bush once stated, "there is too much information" [4]. He proposed a system in which human authors played the role of trailblazers, helping to create pathways through vast networks of information. In the years since publishing those words, the amount of information has exploded. The ability of human authors to traverse huge information collections and organize them in meaningful ways for others has become impractical. In response to problems of this sort, Crane has observed the need for subsystems within digital collections that support two-stage collection development processes that alternate automated and human involvement. His examples include tagging a corpus as part of the editorial process [8], and automatically generating links to external datasets from elements within a large digital library collection [9]. The first stage is conducted automatically by a software agent or other subsystem. The second stage involves refinement by a human. There may be some iterative reformulation. Through the characteristic of alternating automated and human processes, we see corpus tagging and automatic link generation as examples of a more general imperative.

Semi-autonomous agents are needed to assist collection curators in knowledge organization tasks. Unfortunately, computerized systems have yet to discern meaning in the underlying data which they process. Without true understanding, these agents require human guidance and feedback in order to be truly successful. By semi-autonomous agents, we refer to a two-stage process, in which an agent runs, until it reaches some state of partial completeness, and then asks a user for input. Taking this process one step further is to design *mixed-initiative* systems [17], such as the LookOut extension to Microsoft Outlook [17] and combinFormation [22, 23], which integrate the automatic actions of agents with the interactive actions of humans, in order to accomplish complex human centered information processing. Mixed-initiative systems allow for human experts to tailor the underlying models used by the agents to better deal with the semantic subtleties that lie outside the model's capabilities. We need semi-autonomous and mixed initiative systems for developing and utilizing digital collections, which seek to structure the vast amount of information in ways that reduce the cognitive load on the human during processes of classification. The goal is to free valuable human cognitive cycles for processing the real meaning of the uncovered knowledge. The work described in this paper is framed by our broader efforts to build a mixed-initiative system (combinFormation) that assists people in combing through large digital collections, assembling information elements relevant to a task or activity, and forming new ideas about the relationships between these elements.

Pattern recognition techniques strive to discover the structure of information that naturally occurs in multidimensional feature spaces. The features themselves create a perspective through which the underlying structure of the information elements under study can be represented and utilized by agents and system designers. While an agent does not actually understand what it is looking at, it is still able to use the organization of the information encoded through the perspective formulated by human experts. Through the process of gathering training data from experts about what is a junk surrogate, and operating on this knowledge with pattern recognition algorithms, we are able to achieve reasonably effective performance for eliminating junk surrogate candidates. This will enable agents and mixed-initiative systems like combinFormation to more effectively choose surrogates to represent documents, and thus to represent collections of documents as collections of surrogates. Using this pattern recognition method, new semi-autonomous or mixed-initiative

tools may also be developed for authoring a single DSM for each document in a large collection. This, in turn, can further contribute to tools that represent collections of documents as collections of surrogates.

In conclusion, semi-autonomous agents and mixed-initiative systems are necessary for manipulating representations of large digital collections. In surrogate-based systems of this type, pattern recognition based approaches to the elimination of junk surrogate candidates show a great deal of promise. Our initial investigation reveals that the indicated features provide a significant degree of class separability when generalized to a set of HTML documents with diverse structures and styles. Further work can develop the application of better classifiers. Additionally, the DSM structure that we have introduced to define and identify tag pattern features may find broader application in the representation of digital collections.

8. ACKNOWLEDGMENTS

Support is provided by NSF grant IIS-0633906.

9. REFERENCES

[1] Arasu, A., Garcia-Molina, H., Extracting Structured Data from Web Pages, *Proc SIGMOD 2003*, 337-348.

[2] Bishop, C., Neural Networks for Pattern Recognition, Oxford

[3] Burke, M., *Organization of Multimedia Resources*, Hampshire, UK: Gower, 1999.

[4] Bush, V., As We May Think, *The Atlantic Monthly*, July 1945

[5] Chang, C., Lui, S., IEPAD: Information Extraction Based on Pattern Discovery, *Proc of WWW 2001*, 681-688.

[6] Chang, M., Leggett, J., Furuta, R., Kerne, A., Williams, J., Burns, S., Bias, R., Collection Understanding, *Proc JCDL 2004*, 334-342.

[7] cnn.com, viewed 1/27/005

[8] Crane, G., Rydbreg-Cox, J, New Technology and New Roles: The Need for "Corpus Editors", *Proc JCDL 2000*, 252-254

[9] Crane, G., Smith, D.A., Wulfman, C.E., Building a hypertextual digital library in the humanities: a case study on London, *Proc JCDL 2001*, 426-434.

[10] Ding, W., Marchionini, G., Soergel, D., Multimodal Surrogates for Video Browsing, *Proc DL 1999*, 85-93.

[11] Duda R. O., Hart P. E., Stork D. G., 2001, *Pattern Classification, 2nd ed*., Wiley.

[12] Furuta, R. Shipman, F., Marshall, C. Brenner, D., Hsieh, H., Hypertext paths and the World-Wide Web: experiences with Walden's Paths. *Proc ACM Hypertext*, 167-176, 1997.

[13] Glenberg, A.M., Langston, W.E., Comprehension of illustrated text: Pictures help to build mental models, Journal of Memory & Language, 31(2):129-151, April 1992.

[14] Google, http://www.google.com/

[15] Greene, S., Marchionini, G., Plaisant, C., Shneiderman, B., Previews and Overviews in Digital Libraries: Designing Surrogates to Support Visual Information Seeking, *JASIS* 51(4): 380-393, 2000.

[16] Haykin, S., Neural Networks: A Comprehensive Foundation (2nd Ed.), Prentice Hall, 1998.

[17] Horvitz, E., 1999. Principles of Mixed-Initiative User Interfaces, *Proc CHI 1999*, 159-166.

[18] Interface Ecology Lab, *combinFormation*, Texas A&M University:, http://ecologylab.cs.tamu.edu/combinFormation/

[19] Kerne, A., Smith, S.M., The Information Discovery Framework. *Proc ACM Designing Interactive Systems (DIS)* 2004, 357-360.

[20] Kerne, A., Smith, S.M., Choi, H., Graeber, R., Caruso, D., Evaluating Navigational Surrogate Formats with Divergent Browsing Tasks, Proc CHI 2005 Extended, in press.

[21] Kerne, A. Smith S.M., Mistrot, J.M., Sundaram, V., Khandelwal, M., Wang, J., Mapping Interest and Design to Facilitate Creative Process During Mixed-Initiative Information Composition, *Proc Creativity & Cognition Symposium: Interaction: Systems, Practice and Theory*, 2004.

[22] Kerne, A., Sundaram, V., A Recombinant Information Space, *Proc Computational Semiotics in Games and New Media (CoSIGN) 2003*, 48-57.

[23] Kerne, A., Sundaram, V., Wang, J., Khandelwal, M., Mistrot, J.M., Human + Agent: Creating Recombinant Information, *Proc ACM Multimedia 2003*, 454-455.

[24] Lin, S., Ho, J., Discovering Informative Content Blocks from Web Documents, *Proc SIGKDD, 2002*, 588-593.

[25] Marchionini, G., *Information Seeking in Electronic Environments*, Cambridge U Press, 1997.

[26] m.c. schraefel, Shadbolt, N.R., Gibbins, N., Glaser, H., Harris, S., CS AKTive Space: Representing Computer Science in the Semantic Web, *Proc WWW 2004*.

[27] NSDL, The National Science Digital Library, http://nsdl.org/

[28] Payette, S., Lagoze, C., Value-Added Surrogates for Distributed Content, *D-Lib Magazine*, 6:6, June 2000, http://www.dlib.org/dlib/june00/payette/06payette.html

[29] Rowe, N., Coffman, J., Degirmenci, Y., Hall, S., Lee, S., Williams, C., Automatic Removal of Advertising from Web-Page Display, *Proc JCDL, 2002*, 406.

[30] Webb A. R., *Statistical Pattern Recognition*, 2nd ed., Wiley.

[31] Wildemuth, B., Marchionini, G., Yang, M., Geisler, G., Wilkens, T., Hughes, A., Gruss, R., How Fast Is Too Fast? Evaluating Fast Forward Surrogates for Digital Video, *Proc JCDL 2003*, 221-230.

[32] Woodruff, A., Rosenholtz, R., Morrison, J., Faulring, A., Pirolli, P., A Comparison of the Use of Text Summaries, Plain Thumbnails, and Enhanced Thumbnails for Web Search Tasks." JASIST 53(2):172-185, 2002.

[33] W3C, *Cascading Style Sheets*, http://www.w3.org/Style/CSS/

[34] W3C, *Document Object Model (DOM) Level 2 Core Specification*, http://www.w3.org/TR/2000/REC-DOM-Level-2-Core-20001113/, 2000.

[35] Yahoo, http://www.yahoo.com/