Inferring Hierarchical Descriptions

Eric Glover, David M. Pennock, Steve Lawrence, and Robert Krovetz
NEC Research Institute
4 Independence Way
Princeton, NJ, 08540
{compuman,dpennock,lawrence,krovetz}@research.nj.nec.com

ABSTRACT

We create a statistical model for inferring hierarchical term relationships about a topic, given only a small set of example web pages on the topic, without prior knowledge of any hierarchical information. The model can utilize either the full text of the pages in the cluster or the context of links to the pages. To support the model, we use "ground truth" data taken from the category labels in the Open Directory. We show that the model accurately separates terms in the following classes: *self* terms describing the cluster, *parent* terms describing more general concepts, and *child* terms describing specializations of the cluster. For example, for a set of biology pages, sample *parent*, *self*, and *child* terms are *science*, *biology*, and *genetics* respectively. We create an algorithm to predict *parent*, *self*, and *child* terms using the new model, and compare the predictions to the ground truth data. The algorithm accurately ranks a majority of the ground truth terms highly, and identifies additional complementary terms missing in the Open Directory.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Linguistic processing*

General Terms

Experimentation, Measurement, and Algorithms

Keywords

feature selection, hierarchical relationships, statistical models, web analysis, cluster naming

1. INTRODUCTION

Starting with a set of documents, it is desirable to infer automatically various information about that set. Information such as a meaningful name or some related concepts may be useful for searching or analysis. This paper presents a simple model that identifies meaningful classes of features to promote understanding of a cluster of documents. Our simple model defines three types of features: *self* terms that describe the cluster as a whole, *parent* terms that describe more general concepts, and *child* terms that describe specializations of the cluster.

CIKM'02, November 4–9, 2002, McLean, Virginia, USA.

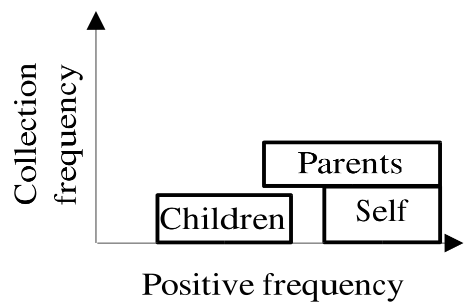

Figure 1: A figure showing the predicted relationships between parent, child and self features. Positive frequency is the percentage of documents in the positive set that contain a given feature. Collection frequency is the overall percentage of documents that contain a given feature.

Automatic selection of parent, child and self features can be useful for several purposes including automatic labeling of web directories or improving information retrieval. An important use could be for automatically naming generated clusters, as well as recommending both more general and more specific concepts, using only the summary statistics of a single cluster, and background collection statistics. Also, popular web directories such as Yahoo (`http://www.yahoo.com/`) or the Open Directory (`http://www.dmoz.org/`) are manually generated and manually maintained. Even if categories are defined by hand, automatic hierarchical descriptions can be useful to recommend new parent or child links, or alternate names. The same technology could be useful to improve information retrieval by recommending alternate queries (both more general and more specific) based on a retrieved set of pages.

1.1 The Model

We hypothesize that we can distinguish between *parent*, *self*, and *child* features based on analysis of the frequency of a feature f in a set of documents (the "positive cluster"), compared to the frequency of f in the entire collection. Specifically, if f is very common in the positive cluster, but relatively rare in the collection, then f may be a good *self* term. A feature that is common in the positive cluster, but also somewhat common in the entire collection, is a description of the positive cluster, but is more general and hence may be a good *parent* feature. Features that are somewhat common in the positive cluster, but very rare in the general collection, may be good *child* features because they only describe a subset of the positive documents.

Figure 1 shows a graphical representation of the model. The three regions define the predicted relative relationships between parent, child and self features. Features outside of the marked regions are considered poor candidates for the classes of parent, child or self. Figure 1 does not show any absolute numerical boundaries, only the relative positions of the regions. The actual regions may be fuzzy or non-rectangular. The regions depend on the generality of the class. For example, for the cluster of "biology" the parent of "science" is relatively common. For a cluster of documents about "gene sequencing", a parent of "DNA" may be more rare than "science", and hence the boundary between parent and self would likely be closer to 0.

Figure 2 shows a view of a set of documents that are in the areas of "science", "biology", and "botany". The outer circle represents the set of all documents in the subject area of "science". The middle circle is the set of documents in the area of "biology" and the inner-most circle represents the documents in the area of "botany". If we assume that the features "science", "biology" and "botany" occur only within their respective circles, and occur in each document contained within their respective circles, it is easy to see the parent, child, self relationships. From this figure, roughly 20% of the total documents mention "science", about 5% of the documents mention "biology" and about 1% mention "botany". Within the set of "biology" documents, 100% mention both "science" and "biology", while about 20% mention "botany". This is a very simplistic representation, because we assume that every document in the biology circle actually contains the word biology – which is not necessarily the case. Likewise, it is unlikely that all documents in the sub-category of botany would mention both "biology" and "science".

To compensate for this, we assume that there is some probability a given "appropriate" feature will be used. This probability is likely less for the parents than for the selfs or children. As a result, in Figure 1, the parent region extends more to the left than the self region. The probability of a given feature being used will also affect the coordinates of the lower right corner; a lower probability may shift the percentage of occurrences in the self to the left. A probability of one would correspond to every positive document containing all self features.

2. AN EXPERIMENT

To test the model described in Figure 1, we used ground truth data and known positive documents to generate a graph of the actual occurrences of parent, self and child features. We chose the Open Directory (`http://www.dmoz.org/`) as our ground truth data for both parent, child and self terms, as well as for the documents. Using the top level categories of "computers", "science" and "sports", we chose

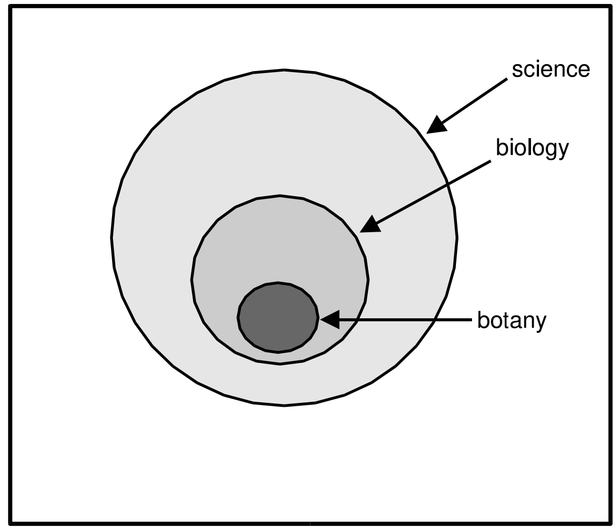

Figure 2: Sample distribution of features for the area of biology, with parent science, and child botany.

the top 15 subject-based sub-categories from each (science only had 11 subject-based sub-categories) for a total of 41 categories to form the set of positive clusters. Table 1 lists the 41 categories, and their parents, used for our experiment. We randomly chose documents from anywhere in the Open Directory to collect an approximation of the collection frequency of features. The negative set frequencies of the parent, children and self features should be similar (between sub-categories) because all 41 sub-categories are at a similar depth (with respect to the Open Directory root node).

Parent	Categories
Science	Agriculture, Anomalies and Alternative Science, Astronomy, Biology, Chemistry, Earth Sciences, Environment, Math, Physics, Social Sciences, Technology
Computers	Artificial Intelligence, CAD, Computer Science, Consultants, Data Communications, Data Formats, Education, Graphics, Hardware, Internet, Multimedia, Programming, Security, Software, Systems
Sports	Baseball, Basketball, Cycling, Equestrian, Football, Golf, Hockey, Martial Arts, Motorsports, Running, Skiing, Soccer, Tennis, Track and Field, Water Sports

Table 1: The 41 Open Directory categories, and the three parent categories we used for our experiment.

Each category has an assigned parent (in this case either science, computers or sports), an associated name, which formed the self features, and several sub-categories, which formed the children. In each case, we split the assigned names on "and", "or", or punctuation such as a comma. So the category of "Anomalies and Alternative Science" becomes two selfs, "anomalies" and "alternative science".

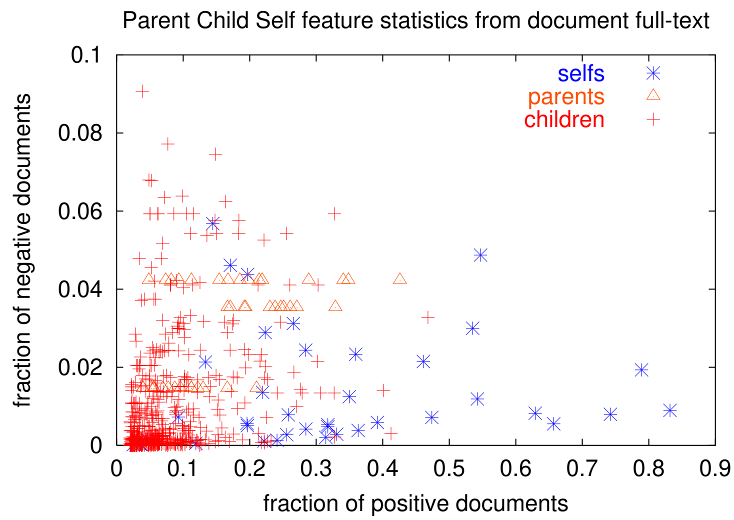

Figure 3: Distribution of ground truth features from the Open Directory.

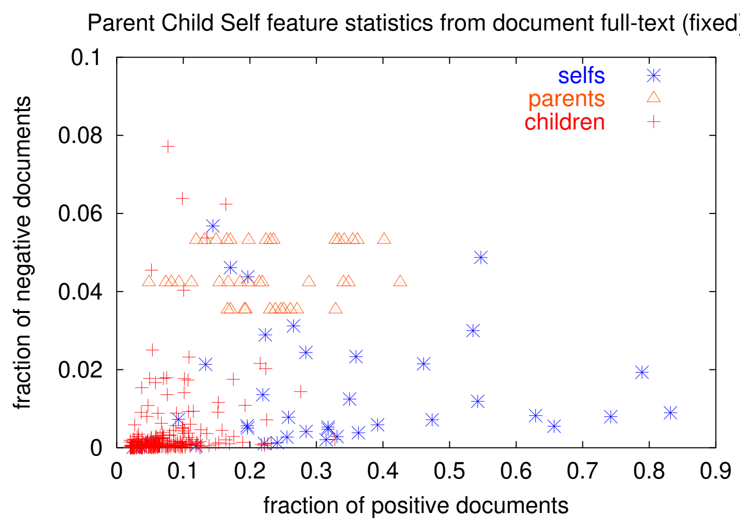

Figure 4: Distribution of ground truth features from the Open Directory, removing the insufficiently defined children, and changing the parent of "computers" to "computer".

The first part of the experiment considered an initial set of 500 random documents from each positive category, and 20,000 random documents from anywhere in the directory as the negative data (collection statistics). Each of the web URLs was downloaded and the features were put into a histogram. If a URL resulted in a terminal error, the page was ignored, explaining the variation in the number of positive documents used for training. Features consisted of words, or two or three word phrases, with each feature counting a maximum of once per document.

Then, for each category, we graphed each parent, child and self feature (as assigned by the Open Directory) with the X coordinate as the fraction of positive documents containing the feature, and the Y coordinate as the fraction of the negative documents containing that feature. If a feature occurred in less than 2% of the positive set it was ignored.

Figure 3 shows the distribution of all parent, child and self features from our 41 categories. Although there appears to be a general trend, there are many children that occur near the parents. Since there were many categories with the same parent (only three unique parents), and a common negative set was used, the parents are co-linear with a common value.

Several of the children are words or phrases that are not well defined in the absence of knowledge of the category. For example, the feature "news" is undefined without knowing the relevant category; is it news about artificial intelligence, or news about baseball? Likewise several features, including news, are not "subjects" but rather a non-textual property of a page. A volunteer went through the list of categories and their children, removing any child that was not sufficiently defined in isolation. He removed more than half of the children. The removal was done prior to seeing any

Category	F	V	Category	F	V
agriculture	438	67	anomalies and alternative science	395	63
artificial intelligence	448	77	astronomy	438	64
baseball	419	62	basketball	418	67
biology	454	66	cad	405	65
chemistry	443	70	computer science	346	75
consultants	442	139	cycling	438	65
data communications	439	65	data formats	434	62
earth sciences	445	70	education	436	67
environment	439	76	equestrian	433	62
football	426	71	golf	441	64
graphics	454	69	hardware	451	67
hockey	411	70	internet	446	74
martial arts	461	61	math	460	69
motorsports	445	64	multimedia	427	64
physics	441	69	programming	446	76
running	436	82	security	426	67
skiing	421	69	soccer	439	73
social sciences	458	71	software	446	73
systems	447	54	technology	439	53
tennis	452	36	track and field	384	60
water sports	451	40			

Table 2: The number of positive documents from each category for the full-text (F) experiment and for the extended anchortext (V) experiment.

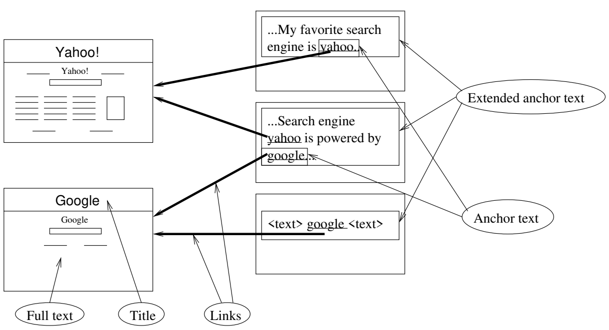

Figure 5: Extended anchortext refers to the words in close proximity to an inbound link.

data, and without knowledge of exactly why he was asked to remove "insufficiently defined" words or phrases.

Analyzing the data suggested that the parent of "computers" should be replaced by "computer". Unlike the word "sports" often found in the plural when used in the general sense, "computers" is often found in the singular form. For this experiment, we did not perform any stemming or stop-word removal, so "computers" and "computer" are different features. Figure 4 shows the same data as Figure 3 except with the parent changed from "computers" to "computer", and the insufficiently defined children removed. This change produces a clearer separation between the regions.

2.1 Extended Anchortext

Unfortunately, documents often do not contain the words that describe their category. In the category of "Multimedia" for example, the feature "multimedia" occurred in only 13% of the positive documents. This is due to a combination of choice of terms by the page authors as well as the fact that often a main web page has no textual contents, and is represented by only a "click here to enter" image.

Our model assumes the "documents" are actually descriptions. Rather than use the words on the page itself, we decided to repeat the experiment using human assigned descriptions of a document in what we call "extended anchortext", as shown in Figure 5. Our earlier work [3] describes extended anchortext, and how it produces features more consistent with the "summary" than the full text of documents. Features found using extended anchortext generated clusters appear to produce more reasonable names.

Extended anchortext refers to the words that occur near a link to the target page. Figure 5 shows an example of extended anchortext. Instead of using the full text, we used a virtual document composed of up to 15 extended anchortexts. Inbound links from Yahoo! or the Open Directory were excluded. When using virtual documents created by considering up to 25 words before, after and including the inbound anchortexts, there is a significant increase in the usage of self features in the positive set (as compared to the full-texts). In the category of Multimedia, the feature "multimedia" occurred in 42% of the positive virtual documents, as opposed to 13% of the full texts. The occurrence of the feature "multimedia" in the negative (random) set was nearly identical for both the full text and the virtual documents, at around 2%.

Table 2 lists the number of positive virtual documents used for each category (randomly picked from the 500 used in the first experiment). We used 743 negative virtual documents as the negative set. However, the generation of virtual documents is quite expensive, forcing us to reduce the total number of pages considered. The improved summarization ability from virtual documents should allow us to operate with fewer total documents.

Figure 6 shows the results for all parents, children and selfs for the extended anchortext. The positive percentages have in general shifted to the right, as selfs become more clearly separated from children. Figure 7 shows the results after removal of the insufficiently defined children and replacing "computers" with "computer". Very few data points fall outside of a simple rectangular region defined around each class. Even including the insufficiently defined children, the three regions are well defined.

Despite the fact that most parents, children, and selfs fall into the shown regions, there are still several factors causing problems. First, we did not perform any stemming. Some features may appear in both singular and plural forms, with one being misclassified. In addition, phrases may occur less often than their individual terms, making selfs appear falsely as children, such as the case of "artificial intelligence", where it appears as a child due to the relatively low occurrence of the phrase.

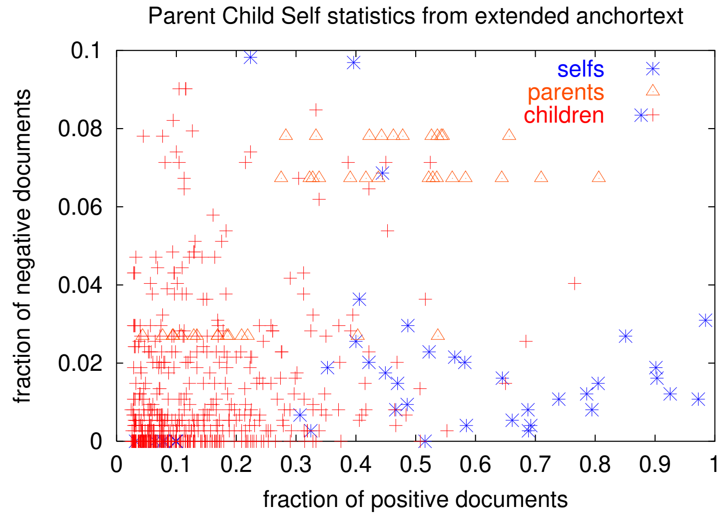

Figure 6: Distribution of ground truth features from the Open Directory using extended anchortext virtual documents instead of full-text.

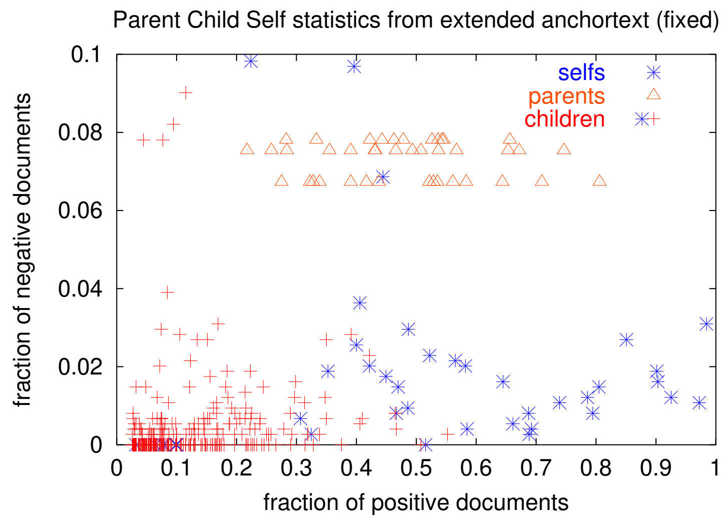

Figure 7: Distribution of ground truth features from the Open Directory using extended anchortext virtual documents instead of full-text, with corrections.

3. EXTRACTING HIERARCHICAL DESCRIPTIONS

3.1 Algorithm

Figure 7 shows that graphing of the ground-truth features from the Open Directory for 41 categories in general follows the predicted model of Figure 1. However, it does not graph all features occurring in each category, only those assigned by The Open Directory. To provide extra support for the model, we present a simple algorithm that ranks all features as possible parents, children and selfs, and compare the output with the ground-truth data from the Open Directory.

Predict Parents, Children and Selfs Algorithm

For each feature f from a set of positive features:

1: Assign a label to feature f as follows:

if $(f.neg > maxParentNegative)$\{Label='N'\}

elseif $(f.neg > maxSelfNegative)$\{Label='P'\}

elseif $(f.pos > minSelfPositive)$\{Label='S'\}

elseif $((f.pos < maxChildPositive)$ and

$(f.neg < maxChildNegative))$\{Label='C'\}

else \{Label='N'\}

2: For each label (P,S,C) sort each feature f with that label by $f.pos$

Category	Parents	Selfs	Children
agriculture	management, science	agriculture, agricultural	soil, sustainable, crop
anomalies and alternative science	articles, science		alternative, ufo, scientific
artificial intelligence	systems, computer	artificial, intelligence	ai, computational, artificial intelligence
astronomy	science, images	space, astronomy	physics, sky, astronomical
baseball	sports, high	baseball, league	stats, players, leagues
basketball	sports, college	basketball, team	s basketball, espn, hoops
biology	science, university of	biology	biological, genetics, plant
cad	systems, computer	cad, 3d	modeling, architectural, 2d
chemistry	science, university of	chemical, chemistry	chem, scientific, of chemistry
computer science	systems, computer	engineering, computing	programming, papers, theory
consultants	systems, management	solutions, consulting	consultants, programming, and web
cycling	sports, url	bike, bicycle	bicycling, mtb, mountain bike
data communications	systems, management	communications, solutions	networks, clients, voice
data formats	collection, which	windows, graphics	file, mac, truetype
earth sciences	science, systems	environmental, data	survey, usgs, ecology
education	computer, training	learning	microsoft, tutorials, certification
environment	science, management	environmental, environment	conservation, sustainable, the environment
equestrian	training, sports	horse, equestrian	riding, the horse, dressage
football	sports, board	football, league	teams, players, leagues
golf	sports, equipment	golf, courses	golfers, golf club, golf course
graphics	images, collection	graphics	3d, animation, animated
hardware	computer, systems	hardware, technologies	hard, components, drives
hockey	sports, canada	hockey, team	hockey league, teams, ice hockey
internet	computer, support		web based, rfc, hosting
martial arts	arts, do	martial, martial arts	fu, defense, kung fu
math	science, university of	math, mathematics	theory, geometry, algebra
motorsports	photos, sports	racing, race	driver, track, speedway
multimedia	media, video	digital, flash	3d, animation, graphic
physics	science, university of	physics	scientific, solar, theory
programming	systems, computer	programming, code	object, documentation, unix
running	sports, training	running, race	races, track, athletic
security	systems, computer	security, system	security and, nt, encryption
skiing	sports, country	ski, skiing	winter, snowboarding, racing
soccer	sports, url	soccer, league	teams, players, leagues
social sciences	science, university of	social	economics, theory, anthropology
software	systems, computer	windows, system	application, tool, programming
systems	computer, systems	computers, hardware	linux, emulator, software and
technology	systems, university of	engineering	scientific, engineers, chemical
tennis	sports, professional	tennis, s tennis	men s, women s tennis, of tennis
track and field	sports, training	running, track	track and field, track and, and field
water sports	board, sports	boat	sailing, boats, race

Table 3: Algorithm predicted top two parents, selfs and children for each of the 41 tested categories. Blank values mean no terms fell into the specified region for that category.

3.2 Results

Using the data from Figure 7, we specified the following cutoffs:

$maxParentNegative = 0.08$

$maxSelfNegative = 0.06$

$minSelfPositive = 0.4$

$maxChildPositive = 0.4$

$maxChildNegative = 0.02$

Table 3 shows the top parents, selfs and children generated using the algorithm described in Section 3.1 as applied to the virtual documents, as described in Section 2.1. The results show that in all 41 categories the Open Directory assigned parent (replacing "computer" for "computers") was ranked in the top 5. In about 80% of the categories the top ranked selfs were identical, or effectively the same (synonym, or identical stem) as the Open Directory assigned self. Children are more difficult to evaluate since there are many reasonable children that are not listed.

Although in general the above algorithm appears to work, there are several obvious limitations. First, in some categories, such as "Internet", the cut-off points vary. Our algorithm does not dynamically adjust to the data for a given category. The manually assigned cut-offs simply show that if we did know the cut-offs the algorithm would work; it does not specify how to obtain such cut-offs automatically. Second, phrases appear to sometimes have a lower positive occurrence than single words. For example, the phrase "artificial intelligence" incorrectly appears as a child instead of a self. Third, there is no stemming or intelligent feature removal. For example, a feature such as "university of" should be ignored since it ends with a stop word. Likewise, consulting as opposed to consult, or computers as opposed to computer are all examples where failure to stem has caused problems.

Despite the problems, the simplistic algorithm suggests that there are some basic relationships between features that can be predicted based solely on their frequency of occurrence in a positive set and in the whole collection. Clearly more work and more detailed experiments are needed.

It should be noted that these categories are all at roughly the same depth (from the root node of the open directory). This increases the likelihood that the cut-offs work for multiple categories, even though each category may be different.

Analysis of the documents in the clusters revealed that some categories suffered from topic drift when random documents were chosen. Our method for choosing the pages for each positive cluster randomly picked pages from the set of all documents in the category or one-level below. Unfortunately, since the Open Directory does not guarantee an equal number of documents in a category, it is possible to pick a higher percentage of documents from one child. For example, in the category of "Multimedia" there are only six URLs in the category itself, with 560 pages in the child of "Flash and Shockwave". Randomly picking documents in that category biases "flash and shockwave" over the more general multimedia pages.

4. RELATED WORK

4.1 Cluster Analysis

There is a large body of related work on automatic summarization. For example, Radev and Fan [9] describe a technique for summarization of a cluster of web documents. Their approach breaks down the documents into individual sentences and identifies themes or "the most salient passages from the selected documents". Their approach uses "centroid-based summarization" and does not produce sets of hierarchically related features.

Lexical techniques have been applied to infer various concept relationships from text [1, 4, 5]. Hearst [5] describes a method for finding lexical relations by identifying a set of lexicosyntactic patterns, such as a comma separated list of noun phrases, e.g. "bruises, wounds, broken bones or other injuries". These patterns are used to suggest types of lexical relationships, for example bruises, wounds and broken bones are all types of injuries. Caraballo describes a technique for automatically constructing a hypernym-labeled noun hierarchy. A hypernym defines a relationship between word A and word B if "native speakers of English accept that sentence B is a (kind of) A". Linguistic relationships such as those described by Hearst and Caraballo are useful for generating thesauri, but do not necessarily describe the relationship of a cluster of documents to the rest of a collection. Knowing that say "baseball is a sport" may be useful for hierarchy generation if you knew a given cluster was about sports. However, the extracted relationships do not necessarily relate to the actual frequency of the concepts in the set. Given a cluster of sports documents that discusses primarily basketball and hockey, the fact that baseball is also a sport is not as important for describing that set as other relationships.

Sanderson and Croft [10] presented a statistical technique based on subsumption relations. In their model, for two terms x and y, x is said to subsume y if the probability of x given y is one,[1] and the probability of y given x is less than one. A subsumption relationship is suggestive of a parent-child relationship (in our case a self-child relationship). This allows a hierarchy to be created in the context of a given cluster. In contrast, our work focuses on specific general regions of features identified as "parents" (more general than the common theme), "selfs" (features that define or describe the cluster as a whole) and "children" (features that describe common sub-concepts). Their work is unable to distinguish between a "parent-self" relationship and a "self-child" relationship. They only deal with a positive set of documents, but statistics from the entire collection are needed to make both distinctions. Considering the collection statistics can also help to filter out less important terms that may not be meaningful to describe the cluster.

Popescul and Ungar describe a simple statistical technique using χ^2 for automatically labeling document clusters [8]. Each (stemmed) feature was assigned a score based on the product of local frequency and predictiveness. Their concept of a good cluster label is similar to our notion of "self features". A good self feature is one that is both common in the positive set and rare in the negative set, which corresponds to high local frequency and a high predictiveness.

Our earlier work [3], describes how ranking features by expected entropy loss can be used to identify good candidates for self names or parent or child concepts. Features that are common in the positive set, and rare in the negative set make good selfs and children, and also demonstrate high expected entropy loss. Parents are also relatively rare in the negative set, and common in the positive set and are also likely to have high expected entropy loss. This work focuses on separating out the different classes of features by considering the specific positive and negative frequencies, as opposed to ranking by a single entropy-based measure.

[1] They actually used 0.8 instead to reduce the noise.

4.2 Hierarchical Clustering

Another approach to analyzing a single cluster is to break it down into sub-clusters — forming a hierarchy of clusters. Fasulo [2] provides a nice summary of a variety of techniques for clustering (and hierarchical clustering) of documents. Kumar et al. [7] analyze the web for communities, using the link structure of the web to determine the clusters. Hofmann and Puzicha [6] describe several statistical models for co-occurrence data and relevant hierarchical clustering algorithms. They specifically address the IR issues and term relationships.

To clarify the difference between our work and hierarchical clustering approaches, we will present a simple example. Imagine a user does a web search for "biology", and retrieves 20 documents, all of them general biology "hub" pages. Each page is somewhat similar in that they don't focus on a specific aspect of biology. Hierarchical clustering would break the 20 documents down into sub-clusters, where each sub-cluster would represent the "children" concepts. The topmost cluster could arguably be considered the "self" cluster. However, given the sub-clusters, there is no easy way to discern which features (words or phrases) are meaningful names. Is "botany" a better name for a sub-cluster than "university"? In addition, given a group of similar documents, the clustering may not be meaningful. The sub-clusters could focus on irrelevant aspects - such as the fact that half of the documents contain the phrase "copyright 2002", while the other half do not. This is especially difficult for web pages that are lacking of textual content, i.e. a "welcome page" or a javaScript redirect, or if some of the pages were about more than one topic (even though the cluster as a whole is primarily about biology).

Using our approach, the set of the 20 documents would be analyzed (considering the web structure to deal with non-descriptive pages), and a histogram summarizing the occurrence of each feature would be generated (individual document frequency would be removed). Comparing the generated histogram to a histogram of all documents (or some larger reference collection), we would find that the "best" name for the cluster is "biology", and that "science" is a term that describes a more general concept. Likewise, we would identify several different "types" of biology, even though no document may actually cluster into the set. For example, "botany", "cell biology", "evolution", etc. Phrases such as "copyright 2002" would be recognized as unimportant because of their frequency in the larger collection. In addition, the use of the web structure (extended anchortext) can significantly improve the ability to name small sets of documents over just the document full text, dealing with the problems of "welcome pages" or redirects.

5. SUMMARY AND FUTURE WORK

This paper presents a simple statistical model that can be used to predict parent, child and self features for a relatively small cluster of documents. Self features can be used as a recommended name for a cluster, while parents and children can be used to "place" the cluster in the space of the larger collection. Parent features suggest a more general concept, while child features suggest concepts that describe a specialization of the self.

To support our model, we performed two different sets of experiments. First, we graphed ground truth data, demonstrating that actual parent, child, and self features generally obey our predicted model. Second, we described and tested a simple algorithm that can predict parent, child and self features given feature histograms. The predicted features often agreed with the ground truth, and may even suggest new interconnections between related categories.

To improve the algorithm, we will be exploring methods for automatically discovering the boundaries of the regions given only the feature histograms for a single cluster. We also intend to handle phrases differently than single word terms, and include various linguistic techniques to improve the selection process.

6. REFERENCES

[1] Sharon A. Caraballo. Automatic construction of a hypernym-labeled noun hierarchy from text. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, 1999.

[2] D. Fasulo. An analysis of recent work on clustering algorithms. Technical report, University of Washington, 1999. Available at: http://citeseer.nj.nec.com/fasulo99analysi.html.

[3] Eric J. Glover, Kostas Tsioutsioulikilis, Steve Lawrence, David M. Pennock, and Gary W. Flake. Using web structure for classifying and describing web pages. In *Proceedings of the 11th WWW Conference*, Hawaii, 2002.

[4] Marti A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the Fourteenth International Conference on Computational Linguistics*, Nantes, France, 1992.

[5] Marti A. Hearst. Automated discovery of WordNet relations. In Christiane Fellbaum, editor, *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

[6] Thomas Hofmann and Jan Puzicha. Statistical models for co-occurrence data. Technical Report AIM-1625, 1998.

[7] S. Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew Tomkins. Trawling the web for emerging cyber-communities. *WWW8 / Computer Networks*, 31(11-16):1481–1493, 1999.

[8] Alexandrin Popescul and Lyle H. Ungar. Automatic labeling of document clusters. Unpublished manuscript, available at: http://citeseer.nj.nec.com/popescul00automatic.html.

[9] Dragomir R. Radev and Weiguo Fan. Automatic summarization of search engine hit lists. In *Proceedings of ACL'2000 Workshop on Recent Advances in Natural Language Processing and Information Retrieval*, Hong Kong, P.R. China, 2000.

[10] Mark Sanderson and Bruce Croft. Deriving concept hierarchies from text. In *22nd ACM SIGIR Conference*, pages 206–213, 1999.