Metadata in a Mediated Agent Architecture

Grace Crowder
crowder@cs.umbc.edu

July 26, 1996

1 Problem Statement

Given a large, dynamic corpus, distributed over several (possibly hundreds or thousands of) physical sites, traditional information retrieval techniques don't seem to scale. Problems for which traditional Information Retrieval (IR) techniques do not readily scale include routing of queries to appropriate sites, and distribution of incoming text to appropriate sites. In [1] there is discussion of three scalability dimensions: data volume, user base, and data diversity. Therefore, a distributed document information system is necessary.

A distributed document information system consists of multiple connected sites managing individual pieces of a large collection of documents. Each site contains some portion of the total document collection with connections to sites managing other portions of the collection.

In a distributed document information system, it is necessary for sites to have knowledge of pieces of the collection that reside at other sites. What this knowledge should be, how it should be generated and managed, and where it should reside are key questions in the design and use of corpus wide knowledge. In analyzing dissemination of collection wide information, Viles [23] says that "complete dissemination of collection wide information is not required to achieve good [retrieval] effectiveness." We are still left with the questions of what the collection wide knowledge should be.

As a proposed solution, the CAFE (Cooperating Agents Find Everything) system makes use of a mediated agent architecture using metadata to describe the data at each site. Brokers use the metadata to direct activity to appropriate agents. For example, queries will be routed to the appropriate information agents by a broker which uses the metadata supplied by the agents to route the query to the agents whose metadata says a match is probable.

The following presents the CAFE system, its approach to metadata, and the use of the Telltale information retrieval engine to test our hypotheses:

- N-gram profiles can reliably route queries to the right corpus.
- N-gram profiles can reliably route queries to the right portions of a heterogeneous corpus.
- reduced N-gram profiles are also effective, but which reduction techniques work best?

2 Literature Survey

In this survey, we present literature pertinent to the problem stated and solution proposed in Section 1. We are interested in large scale document information systems that are distributed and use metadata to direct agent search of the information.

Topics to be covered include a brief description of information retrieval literature highlighting the use of n-grams as terms, a more detailed discussion of evaluation measures in use in information retrieval, and a look at existing information retrieval systems with an emphasis on distributed large scale document information systems and how they handle the problem of query routing. It is with respect to query routing (for now) that the topic of metadata will be discussed. In addition, KQML will be discussed as that is the language used by CAFE agents.

2.1 Information Retrieval

An information retrieval system is used to manage documents. Required operations include the ability to add, change, delete, display, and query the document corpus [12]. Conceptual models of information retrieval systems are based on boolean retrieval operations, vector space model, probabilistic retrieval, or some combination of these.

In an IR system based on boolean retrieval operations, documents are matched with queries (and thereby considered as answers) when the document contains the specified set of keywords from the query. Boolean queries contain keywords joined by Boolean operators: AND, OR, NOT. In an IR system based on the vector space model, documents and/or queries are considered to be a vector of terms. An answer to a query is a list of documents that contain information relevant to the poser of the query. In an IR system based on the vector space model, an answer to a query is found by computing the cosine of the angle between the vector representing the terms found in the query and the vector representing the terms found in the document. This computation is performed for each document in the corpus and the results are used to determine which documents are most similar to the query. Further discussion of IR models can be found in [12, 19, 25].

Characteristics of Information Retrieval systems that are of interest include: is the system networked? does the system use metadata and what type of metadata is used? what are the terms in the system? what model is used for matching documents to queries?

SMART is an information retrieval system based on the vector space model using words as terms. SMART was developed at Cornell under the direction of Gerard Salton [19]. SMART is a language-dependent IR system in that it performs stop-listing and stemming on input. SMART is not networked and does not use metadata. SMART is a successful participant in the Text Retrieval Conference (TREC) managed by Harman [13] at NIST. Other similar word-based systems that are not networked and do not use metadata include: mg, developed to implement and test the algorithms described in [25], and Inquery, developed by the Center for Intelligent Information Retrieval at the University of Massachusetts at Amherst.

Telltale is an information retrieval system based on the vector space model using n-grams as terms. Telltale is not networked and does not make use of metadata. Telltale was developed at UMBC by Claudia Pearce [17]. As first introduced in Shannon [21], an n-gram is a sequence of n characters. Example 5-grams taken from this sentence: examp, xampl, ample. Pearce [17, 18], Church [6], and Damashek [11] have shown that IR systems based on n-grams as terms have many desirable properties:

- Degraded text is not a problem. As Pearce demonstrated in [17], "the similarity link retrieves documents with similar vocabulary and content to a sample that may contain up to 30 percent random degradation."

Spelling correction algorithms are only effective for regularly degraded text. This means that in a word based information retrieval system, word matches suffer when the document or query text is degraded.

- Multi-lingual text can be handled. See [11] for a discussion of using n-grams to categorize documents based on language. See also [6] for examples using n-grams to find related documents in French and English translations of the same corpus.

  In a word based information retrieval system, language dependence is introduced through two common practices: stemming and stop listing the input. Parsing the input in a word based information retrieval system typically requires stemming algorithms in which prefixes and suffixes are stripped off. In addition it is common to have a list of stop words which are discarded on input. Stop words are words which are so common as to add no discriminating information. Performing stemming and stop listing requires significant re-programming of the information retrieval system for each language to be considered. Use of n-grams as terms avoids this added programming requirement when changing the language of the data.

- Phrases are captured because the interword spaces are part of an n-gram containing the surrounding characters. When the list of n-grams is weighted by occurrence, spaces between words that mostly occur together (i.e, phrases) will be weighted more heavily than spaces that occur between unrelated words. See [7] for a discussion of highlights extracted from documents using n-grams. Highlights are meant to distinguish a document from the background collection of documents. Frequently highlight lists will include phrases.

  In order for word based information retrieval to recognize phrases, significant natural language processing must be performed.

Networked information retrieval systems include:

- Harvest: servers use glimpse to build metadata and send it to a broker when asked [2]. Metadata consists of content summaries extracted by the Essence system. Essence is customizable and applies extraction techniques based on file type.

- WHOIS++: servers build metadata, using templates, that is collected into centroids at indexing servers [24]. WHOIS++ has the notion of abstractable metadata in that a collection of centroids can be stuck together and treated as a centroid for the index servers.

- Lycos: searches http, ftp, and gopher sites. For each file the metadata consists of title, headings and sub-headings, 100 most significant words according to standard tf*idf measure, first 20 lines, size in bytes of the file, and the number of words in the file. This metadata is used to produce a list of possible matches when Lycos does a search.

2.2 Evaluation Measures

Evaluation measures are needed to determine whether the metadata chosen meet the criteria for metadata. We require that metadata have the following properties:

- Effective - if the metadata says the information is [not] relevant, it's probably [not] relevant

- Concise - much smaller than the data it describes

- Generated automatically - little or no human intervention required

- Abstractable - can create metadata of metadata

- Interchangeable - the form must allow queries, documents, and metadata to be compared.

Therefore, evaluation measures for each of these criteria will be applied.

2.2.1 Measures of effectiveness

Two traditional measures of effectiveness in information retrieval are precision and recall. An information retrieval system, such as SMART or Telltale, returns a list of documents in response to a query. In a scored collection, such as the Cranfield collection, every document has been judged by a topic expert as either relevant or not relevant for every query.

Recall measures the percent of relevant documents that the system being tested retrieves in response to a scored query. Recall varies from 0.0 to

1.0, and is the ratio of relevant documents retrieved to the total number of relevant documents in the collection.

Precision measures the percent of documents that the system retrieves that are relevant. Precision also varies from 0.0 to 1.0 and is the ratio of relevant documents retrieved to the total number of documents retrieved. [20, 25]

Precision is usually displayed as a function of recall. As recall improves there is generally a drop in precision.

R-values are used to compare the retrieval effectiveness of two IR systems. The R-value for a query is the computed precision at the point where all relevant documents could have been returned had recall been 100 percent. For example, if there are 29 documents judged relevant for a particular query, then the R-value for that query is the precision value for the 29-th document returned by that system.

Precision and recall are measures that do not take into account the size of the corpus being considered. Fallout is a measure which considers the collection size. Witten, et al. [25] suggest that rather than plot precision as a function of recall, one may want to plot the detection probability (recall) as a function of the probability of a false alarm (fallout). Fallout is the ratio of nonrelevant items retrieved to the number of nonrelevant items in the collection.

2.2.2 Measures of conciseness

An appropriate measure of conciseness will be the ratio of metadata size to corpus size. This will be a measure of resources used to process, send and save metadata. We expect that as the metadata is made more concise effectiveness measures will be reduced. It will be interesting to understand this relationship as it will describe how reducing the metadata will reduce its effectiveness.

2.2.3 Measures of automatic generation

Measure the amount of human input required to compute metadata. This number should be 0 or close to it.

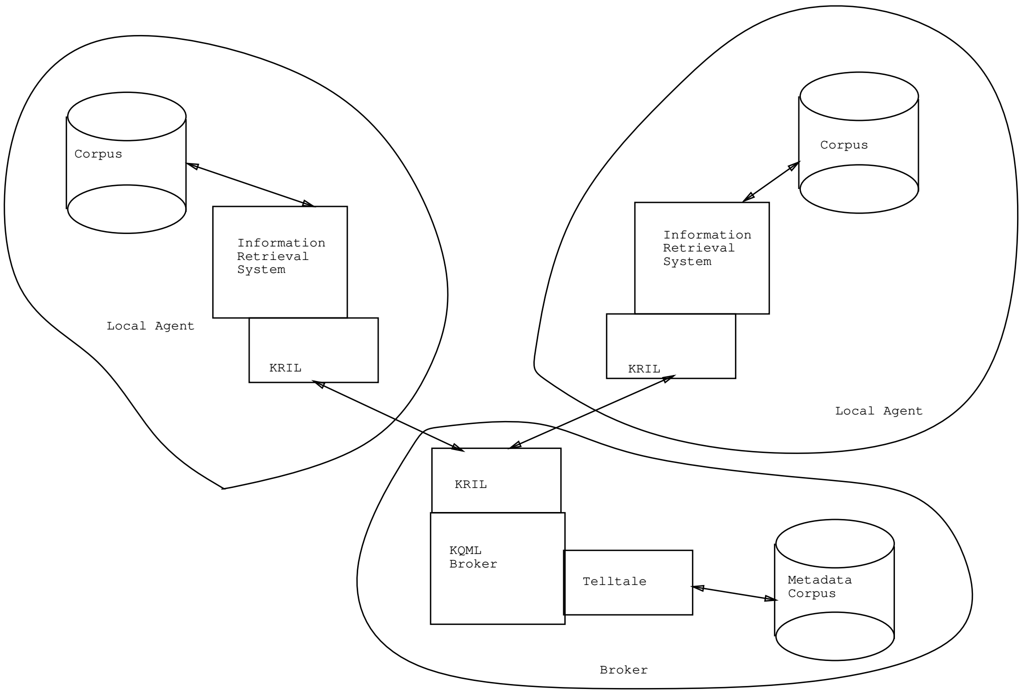

Figure 1: CAFE: Agent Mediated Architecture

2.2.4 Measures of abstractability

Measure the ability to abstract from metadata corpora metadata to be used in a hierarchy of brokers. This will be either possible (at some computational cost) or not possible.

2.2.5 Measures of interchangeability

Interchangeability is achieved when we can apply the same algorithms to queries, documents, corpora, metadata, and collections of metadata.

3 The CAFE System

CAFE (Cooperating Agents Find Everything), see Figure 1, is a large scale document information system consisting of agents managing local corpora using information retrieval engines as backends, brokers using metadata to direct information search, KQML for agent communication, and a user interface to be specified later. KQML is a language designed to support interaction among intelligent software agents [15]. Intelligent agents and brokers communicate to carry out information system needs.

3.1 Agents

In CAFE, a local agent is an information retrieval system that is capable of registering its metadata with an assigned broker, responding to information requests from a broker, and updating its corpus and corresponding metadata. To date, we have been thinking that an agent can be any information retrieval system that can accomplish these tasks. Agents will communicate with assigned brokers using KQML.

3.2 KQML

In CAFE, a subset of KQML performatives are used. Agents register their metadata using the advertise performative, information requests are sent by brokers to agents using the ask performative. An interesting aspect of the use of KQML by CAFE is in using an information system (Telltale) as the agent locator in the broker. In addition, the CAFE broker will be able to merge results of a query sent to multiple agents before presenting those results to an information requestor.

3.3 Brokers

In CAFE, a broker is a KQML type of broker: one capable of registering agents and their metadata, and using that metadata to serve information needs of users. The broker is responsible for determining which agents will be asked for information and for merging the results of the agents' responses before presenting the answer to the user.

3.4 Telltale

Telltale is an information retrieval system based on the vector space model of information retrieval using n-grams as terms [18]. By using Telltale as the information system in the local agents, CAFE provides a mechanism for scaling Telltale to handle very large corpora. Telltale is also used as the information system in the CAFE broker, with metadata as the corpus, to direct information requests to appropriate agents.

4 Metadata

The keys to scalability in the CAFE architecture are the use of a mediated agent architecture as described in [15] and the use of metadata to direct the agent search. This section will define metadata and give the desiderata for metadata in the CAFE system.

Metadata is:

- data about the data. Metadata must describe the data sufficiently well as to be used as a surrogate for the data when making decisions regarding use of the data. In a document information retrieval system, data is added to the system, deleted from the system, retrieved from the system in response to queries, etc. To be useful, metadata must be effective, concise, generated automatically, abstractable, and interchangeable.

- effective if when the metadata says the data is relevant, then the data probably is relevant. Conversely, metadata is effective if when the metadata says the data is not relevant, then the data probably is not relevant. That is, if the metadata can be used in place of the data for a particular task, then the metadata is effective for that task.

- concise for a document information system when it is smaller than the data it represents. In our view of data and metadata, this is a requirement otherwise the data itself would be used. We are interested in scalable solutions and the metadata must be smaller than the data for the system to scale.

- generated automatically by computer based tools. The metadata can be generated as needed and updated at appropriate times. The metadata we are considering and its automatic generation will be described in detail in later sections of this paper.

- abstractable when there can be metadata of metadata. The ability to create metadata from a collection of data that is itself metadata further supports a scalable solution.

- interchangeable when the representation of queries, documents, corpora, and metadata allows for them to be used interchangeably.

Possible metadata includes: the data itself, some natural language description of the data, a template, or something else, such as an n-gram profile.

The data itself is not a reasonable candidate for metadata in the CAFE system because it does not meet all of the requirements for metadata. For example, we require the system scale to handle very large document collections and this would lead to passing large amounts of data as metadata between agents and brokers. Automatic generation of a natural language description of the data requires understanding of the text documents, which is not possible given current tools. Templates, such as USMARC records, as metadata fall short of the requirements that metadata be abstractable and automatically generated.

4.1 N-gram Profiles as Metadata

A frequency distribution of the n-grams in a document is called its n-gram profile. Pearce and Nicholas [18] showed that a document's n-gram profile serves to characterize its content. The mean of the collected n-gram profiles of a set of documents is the centroid. Note that the centroid document isn't a real document. The whole corpus has a centroid as does any subset of member documents.

N-gram profiles:

- are appealing as metadata because they meet the criteria listed above for metadata. They are effective, concise, automatically generated, abstractable, and interchangeable.

- are effective as shown by Pearce and Nicholas using the Telltale system to create hypertext links between documents [18].

- are concise. For n of 5, the n-gram profile for a 40 megabyte Wall Street Journal corpus requires about 6.5 megabytes of storage. See Figure 3.

- can be generated automatically using Telltale [17, 18].

- are abstractable. A collection of n-gram profiles also has an n-gram profile that fits the criteria set forth for metadata. That is, there is no distinction between an n-gram profile for a single document and an n-gram profile for a collection of documents. There is also no distinction between an n-gram profile for a document collection and an n-gram profile for a collection of document collections. As an aside, it is interesting to note that queries in the Telltale information retrieval system are treated by manipulating their n-gram profiles, thus further blurring the traditional distinctions made between query, document, corpus, data, and metadata.

- are interchangeable. A query can be compared to an n-gram profile in the same way that a query can be compared to a document. This allows the existing n-gram engine to be used with metadata as it is used with the whole data.

One caveat in dealing with n-gram profiles is that they can become quite large. In spite of being concise, a 5-gram profile requires about 15 percent of the corpus size for storage. Therefore, we are considering several approaches to reducing the size of the n-gram profile used as metadata.

4.1.1 Approaches to Generating Metadata

A research question to be investigated is what is a reasonable compression ratio for the metadata? That is, how much of the total data can be discarded before the remaining data is not effective? In particular, can metadata be used by the Telltale enabled broker to route queries where data is scattered over numerous nodes and how small can the metadata be made and still fit the criteria?

Approaches to generating metadata include the following.

- Determine upper and lower thresholds to discard n-grams with diminished resolving power from the metadata [22]. It has been observed [8] that n-grams obey many of the same statistical properties that words obey. In particular, n-grams obey Ziph's Law, which states that the product of rank times frequency of words (n-grams) remain constant throughout a work. Therefore, those n-grams that occur so frequently as to add no information about the corpus content can be discarded from the reduced n-gram profile. Likewise, those n-grams that occur so infrequently as to add no information about the corpus content can be discarded from the reduced n-gram profile. The problem is to determine the frequency cutoff points.

- Vary the size of n. As demonstrated by some preliminary experiments, as n grows, so does the size of the n-gram profile. The effectiveness of larger n does not yet seem confirmed. Therefore, in the interests of producing the most concise, effective metadata, choose a small n.

- Vary the window slide factor. In computing the n-grams in a document, we use a sliding window that passes over all of the text processing each n character sequence. This window slides forward one character at a time, thus processing overlapping n-grams [8]. This gives Telltale good resistance to degraded text, but it also means that n-gram profiles contain a lot of redundant information [8, 17]. In this approach the slide factor is one. An increase in the slide factor will reduce the amount of redundant information in an n-gram vector. Choices to consider include 1, 2, $n/2$, n, $2n$.

  In addition, the choice of n and the window slide factor may be dependent on:

  - corpus' average word size
  - the data source. In particular, how 'clean' is the data?
  - each other.

- Mayfield's coverage algorithm [16]. The purpose of this algorithm is to produce a list of n-grams that cover the most text in a document.

- Highlights as discussed by Cohen in [7]. A list of highlights generated from a document is a list of those words and phrases that occur more than the expected number of times in this document given the frequencies observed in the rest of the corpus.

It is important that each approach meet the basic criteria for metadata: it must be effective, concise, automatically generated, and abstractable. In

Collection	Subject	number of documents	number of queries	size in megabytes
LISA	Library Science	5,872	35	3.4
NPL	Electrical Engineering	11,429	93	3.1
CACM	Computer Science	3,204	64	2.2
CISI	Library Science	1,460	112	2.2
Cranfield	Aeronautics	1,400	225	1.6
Time	General Articles	423	83	1.5
Medline	Medicine	1,033	30	1.1
ADI	Information Science	82	35	0.04

Figure 2: Scored Collections

addition, metrics to be used to compare one approach with another must be applied. Metrics such as R-value [25] are useful for comparing two information retrieval systems.

5 Research Plan

I plan to generate the reduced n-gram profile using each of the methods described above and then measure the tradeoffs between conciseness and effectiveness for information retrieval tasks within the CAFE architecture using the reduced n-gram profile as the metadata.

5.1 Experimental Design

Retrieval effectiveness measures require relevance judgements be made prior to exercising the information system being tested. To that end, several test collections are available for use. A test collection is a collection of documents, queries, and relevance judgements. For each query in the collection, each document has been judged by human experts as either relevant or not relevant. Precision and recall measures are used in judging an information retrieval system's performance against a test collection. See a listing of freely available test collections in Figure 2.

One of the problems with using these test collections to gather effectiveness measures for a large scale document information system is that these collections are small. None of them is several gigabytes, or even one gigabyte. In addition, these collections are each homogeneous: they cover only a single topic or writing style. Therefore, our experiments are somewhat contrived. Perhaps, once the experimental methods are in place, a larger corpus will prove useful.

In an effort to use the data available, we have designed two experiments to test our approaches to metadata in a mediated agent architecture. For the experiments to be successful, a KQML broker will be built which uses a modified Telltale with the ability to read in centroids (or whatever metadata scheme is being tested) as input data (as documents are in the plain information system version of Telltale) and do routing by returning centroid names to be converted to KQML agent names. In addition, Telltale-KQML agents will be built and each of the methods for generating a reduced n-gram profile will be implemented in Telltale.

In experiment one we will put each test collection on a single server, have each server compute and register metadata which is used by the broker to route queries, then measure query hits where a hit is defined as the query being routed to the server handling the associated corpus. We expect each query to be routed to the IR system handling the collection it came from. Our hypothesis is that the precision and recall for each collection's set of queries will be the same as if the queries were being handled by a single IR system.

In the second experiment we will shuffle the available corpora by randomly assigning documents to agents. The agents will generate and register their metadata first using the whole n-gram profile and then by using each of the reduction techniques. We will again run all of the queries and record the agents to which queries are routed and the documents returned as responses to the queries. A hit will be defined as a query being routed to an agent managing a document considered relevant to the query. A hit could also be defined in the traditional sense of a relevant document being returned as a response to a query. Our hypothesis is that the metadata will guide the query to the agent with the most relevant documents for answering that query.

We are interested in analyzing the relationship between the reduction techniques. What is the most effective reduction technique? That is, what is the best technique for generating the least amount of metadata and getting the most precision and recall when querying the CAFE system? Perhaps a hybrid reduction technique will be most successful. We say this because in his

N-gram Size	Unique N-grams
2	707
3	10,029
4	65,976
5	244,958
6	681,282
7	1,492,538
8	2,690,108

Figure 3: Number of Unique N-grams in a 40MB WSJ Corpus as a Function of N-gram size

study of threshold settings (how many of the most frequent 3-grams to clip) for overlapping (window slide factor is one) and non-overlapping (window slide factor is three) 3-grams, Comlekoglu [8] found that: "when the threshold reaches 200, the overlapping 3-gram based retrieval system misses only 35 queries out of 1500 – as opposed to the nonoverlapping 3-gram retrieval system which misses 223 out of 1500 submitted." Space savings were also interesting: in the overlapping 3-gram based retrieval system, a threshold of 200 resulted in a 40.31% savings in secondary storage usage and in the nonoverlapping 3-gram based retrieval system, there was a savings of 26.47%.

6 Preliminary Results

We have run some preliminary experiments to examine the role the size of n plays in retrieval effectiveness. This experiment was done using two different corpora. One corpus, a 40 megabyte corpus of the *Wall Street Journal*, is considered a heterogeneous corpus and the second corpus, a 7 megabyte corpus of the *Computists Communique*, is considered a homogeneous corpus. In addition, we have examined Telltale's performance using the Cranfield collection measuring precision and recall.

We have begun some comparisons of Telltale with SMART to examine strengths and weaknesses of the two approaches.

6.1 Size of N

The size of an n-gram profile depends on the choice of n. As n grows, so does the number of n-grams in the n-gram profile. See Figure 3, and discussion in [9], for *The Wall Street Journal* corpus n-gram size vs. number of n-grams.

In an unscored corpus using queries that are segments of documents in the corpus and measuring retrieval effectiveness as numbers of documents returned, the choice of n seems related to heterogeneity of the corpus – for a homogeneous corpus, larger n seems to give better retrieval effectiveness. For a heterogeneous corpus, the number of documents retrieved in response to a fixed set of queries does not seem to be a linear function of n. Further experiments are necessary to determine optimal values of n in relation to other corpus characteristics.

A small experiment was conducted using two different corpora: *The Wall Street Journal* (WSJ) corpus and a corpus made up of collected back issues of *The Computists Communique* (TCC). The WSJ corpus is considered a heterogeneous corpus both in subject matter and in authorship. The TCC corpus is considered a homogeneous corpus both in subject matter and in authorship. For each corpus, a set of 500 queries was constructed using randomly selected pieces of text from the corpus. The number of documents retrieved for each query against its corpus was recorded as n varied from 3 to 6. The resulting figures were then subjected to regression analysis with n as the independent variable and the number of documents retrieved as the dependent variable. In the TCC corpus, the number of documents retrieved seems to depend on the size of n for similarity measures less than or equal to 55%. In the WSJ corpus, the size of n does not seem to be a significant factor in the number of documents returned in response to the queries.

From the experimental results, we conclude that in a heterogeneous corpus, there is no gain in storing large n-grams as metadata. Therefore, in the interest of scalability, the choice of n for corpus wide metadata can be small. Experiments are on-going to determine the relationship between effective n and average word size for a corpus and how this might effect appropriate choices for n.

6.2 Comparison of Telltale and Smart

The use of the Cranfield collection for comparison of Telltale and SMART is discussed in [9]. In comparing Telltale and SMART using the Cranfield collection, we found the r-precision difference of SMART and Telltale averages 0.06 in a possible range of -1.0 to +1.0 [9]. Further use of this scored collection to compare use of metadata centroids will be done in a similar fashion.

7 Interesting Questions

The questions I propose to investigate are:

What is the relationship between:

- reduction techniques? In particular, is there some hybrid reduction technique that is most successful?
- Telltale similarity threshold as set in the broker and reduction techniques used by the agents to generate metadata?
- reduction techniques and corpus characteristics, such as heterogeneity, average word size, etc?

7.1 Research Plan

A KQML broker will be built that accepts metadata from registered agents and uses that metadata to direct searches to appropriate agents. Telltale has been modified to accept metadata as incoming documents with agent identifiers rather than document identifiers being returned for similarity matches. This Telltale will be used by the KQML broker to choose appropriate agents.

In addition, the different approaches to metadata will be tested by performing the experiments described above to measure retrieval effectiveness in traditional IR using Telltale with metadata in place of the complete corpus centroid. These experiments will then be tested on the full CAFE architecture and measures of effectiveness, as well as conciseness, automatic generation, abstractability, and interchangeeability will be made and compared.

8 Outline of Thesis

- Introduction and Problem Statement
- Literature Survey
- Proposed Solution
  - CAFE Architecture - a broker that deals with metadata
  - Metadata Properties
- Reduction Techniques
- Experiments
- Discussion
- Conclusion

References

[1] C. Mic Bowman, Peter Danzig, Udi Manber, and Michael Schwartz. Scalable Internet Resource Discovery: Research Problems and Approaches. *Communications of the ACM*, Vol. 37, no. 8, pp. 98-107, 114(Aug 1994).

[2] C. Mic Bowman, Peter B. Danzig, Darren R. Hardy, Udi Manber, Michael F. Schwartz, and Duane P. Wessels. *Harvest: A Scalable, Customizable Discovery and Access System.* Technical Report CU-CS-732-94, Department of Computer Science, University of Colorado - Boulder. March 1995.

[3] Francis Bretherton. *Reference Model for Metadata: A Strawman.* IEEE Metadata Workshop, May 1994.

[4] CAFE project meetings held at UMBC spring 1996.

[5] William Cavnar. Using an N-Gram-Based Document Representation with a Vector Processing Retrieval Model. *TREC 3*.

[6] Kenneth Ward Church. *Ngrams* Talk given at UMBC, Fall 1995.

[7] Jonathan Cohen. Highlights: Language and Domain Independent Automatic Indexing Terms for Abstracting. *Journal of the American Society for Information Science, (JASIS)*, April 1995.

[8] Fatih Mehmet Comlekoglu (Eryolu). *Optimizing a Text Retrieval System Utilizing N-gram Indexing.* Ph. D. thesis, The George Washington University, 1990.

[9] Grace Crowder and Charles Nicholas. Using Statistical Properties of Text to Create Metadata. *First IEEE Metadata Conference.* April 1996.

[10] Grace Crowder and Charles Nicholas. An Approach to Large Scale Distributed Information Systems Using Statistical Properties of Text to Guide Agent Search. *CIKM Workshop on Intelligent Information Agents*, 1995.

[11] Marc Damashek. Gauging Similarity with n-Grams: Language-Independent Categorization of Text. *Science*, Vol. 267, pp. 843-848, 10 February 1995.

[12] William Frakes and Ricardo Baeza-Yates, eds. *Information Retrieval: Data Structures and Algorithms.* Prentice-Hall, 1992.

[13] Donna Harman. The Third Text REtrieval Conference. NIST Special Publication, PB95-216883, 1995.

[14] Douglas B. Lenat, Ramanathan V. Guha, Karen Pittman, Dexter Pratt, and Mary Sheperd. Cyc: Toward Programs with Common Sense. *Communications of the ACM*, Vol. 33, no. 8, pp. 30-49 (Aug 1990).

[15] James Mayfield, Yannis Labrou, and Tim Finin. Evaluation of KQML as an Agent Communication Language. in M. Wooldridge, J. P. Muller, and M. Tambe(Ed.), *Intelligent Agents II: Agent Theories, Architectures, and Languages*, Springer-Verlag Lecture Notes in AI - Volume 1037, 1996.

[16] James Mayfield. mayfield/hack/ngram/doc.c. Summer, 1995.

[17] Claudia Pearce. *A Dynamic Hypertext Environment Through n-gram Analysis*. Ph. D. thesis, University of Maryland Baltimore County, 1994.

[18] Claudia Pearce and Charles Nicholas. TELLTALE: Experiments in a Dynamic Hypertext Environment for Degraded and Multilingual Data. *Journal of the American Society for Information Science(JASIS)*, April 1996.

[19] Gerard Salton and Michael J. McGill. *Introduction to Modern Information Retrieval*. McGraw Hill, New York, 1983.

[20] Gerard Salton. *Automatic Text Processing: the Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley, 1989.

[21] Claude Shannon. A Mathematical Theory of Communication. *The Bell System Technical Journal*, Vol 27, pp. 379-423, 623-656, July, October 1948.

[22] C.J. van Rijsbergen. *Information Retrieval*, second edition. Butterworths, London, 1979.

[23] Charles L. Viles and James C. French. Dissemination of Collection Wide Information in a Distributed Information Retrieval System. *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* , July, 1995.

[24] Chris Weider, Jim Fulton, Simon Spero. *Architecture of the Whois++ Index Service*, Internet-DRAFT. June, 1995.

[25] Ian H. Witten, Alistair Moffat, Timothy C. Bell. *Managing Gigabytes: Compressing and Indexing Documents and Images*. Van Nostrand Reinhold, New York, 1994.