An Approach to Large Scale Distributed Information Systems Using Statistical Properties of Text to Guide Agent Search

Grace Crowder *
crowder@cs.umbc.edu

Charles Nicholas
nicholas@cs.umbc.edu

Computer Science and Electrical Engineering Department
University of Maryland Baltimore County
Baltimore, MD 21228-5398
September 22, 1995

1 Introduction and 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 [BDMS94]. Web-walking robots are fine as far as they go, but they're centralized, they don't stay current, and, except for Harvest [BDHMSW95], none, as yet, have any notion of local indices that are shared.

2 A Proposed Solution

The solution we propose uses a mediated agent architecture composed of local server agents managing local corpora and communicating via agent name servers and brokers [see Figure 1]. Use of this intelligent agent architecture leverages the scalability solutions from the intelligent agent community for use in distributed information retrieval.

In addition to the use of a mediated architecture, our solution depends upon the use of automatically generated, effective metadata. In this case, metadata is effective when it provides accurate pointers to agents who maintain applicable data. Each local server generates a statistical profile of its text data using our n-gram based text profiler. The statistical profile of the data is registered with the broker agent as metadata. Metadata is used to direct the query, search, and

*This work was supported in part by the United States Department of Defense.

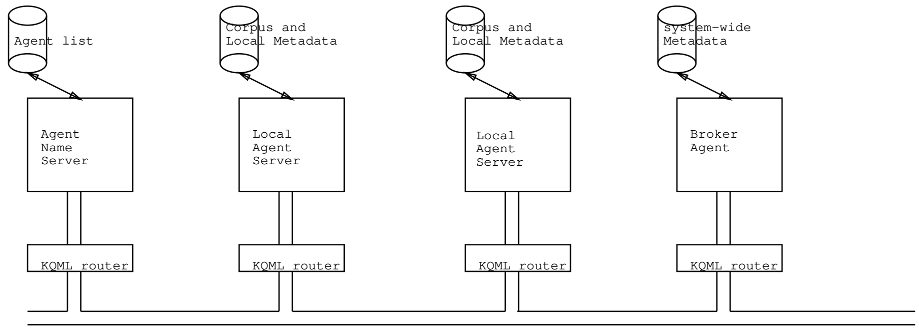

Figure 1: Mediated Architecture for Intelligent Information System

browse operations over the entire (system-wide) corpus. Besides using metadata to direct queries against the system-wide corpus, metadata can also be used to direct incoming data to the appropriate agent according to the statistical profile of the incoming data and the profiles of the agents in the information system. The storing, archiving, and transferring of data can be metadata directed as well.

2.1 Architecture

The information system is a mediated architecture consisting of: local agent servers, with local corpora; agent name servers; and broker agents.

2.1.1 Local Agent Server

A local agent server is responsible for its own data (acts as a server), communicates with its peers using services provided by the agent name server and the broker agents, and knows about its own data as demonstrated through use and maintenance of metadata. By responsible for its own data, we mean the presence of some (arbitrary) information retrieval or database back-end. Note that local servers may "share" their corpus with other services, such as WWW or WAIS servers. See Figure 2.

2.1.2 Agent Name Server

The agent name server maintains a list of agents that have registered. This list is used to facilitate communication between agents. The broker agents maintain metadata registered by the agents. This metadata can be used to direct queries to appropriate agents. We envision agents communicating with each other using Knowledge Query and Manipulation Language (KQML) [FLM95].

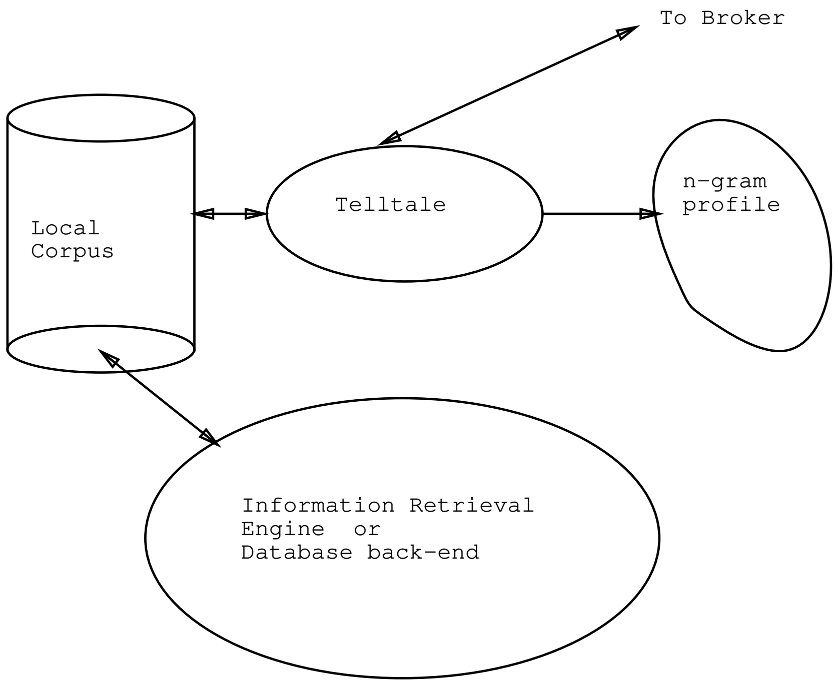

Figure 2: Local Server Agent

2.2 Metadata

Metadata is the key to the scalabiliity of intelligent information agents. Criteria for metadata is that it be automatically generated and effective when used in query, search, browse, and retrieval of the data it describes. We also assume that the metadata is smaller than the data it describes. In addition, metadata can be used to direct incoming data to an appropriate agent, facilitate application to application transfer of data, and manage storing and archiving of data [Bre94]. We will focus on metadata for text. Local agent servers are responsible for generating, updating, and sharing metadata with their designated agent broker(s).

One appealing form of metadata for a corpus is that corpus' n-gram profile. An n-gram profile of a corpus is a list of all n character sequences that occur in the corpus along with their frequency counts [Pea94, PN95, Dam95, Coh95]. The n character sequence is called an n-gram [Sha48]. n-gram based IR is attractive for several reasons, the chief reason being the fact that the n-gram profile of a corpus fits the criteria listed above for metadata. We use KQML for transmission of metadata, the content being n-gram profiles.

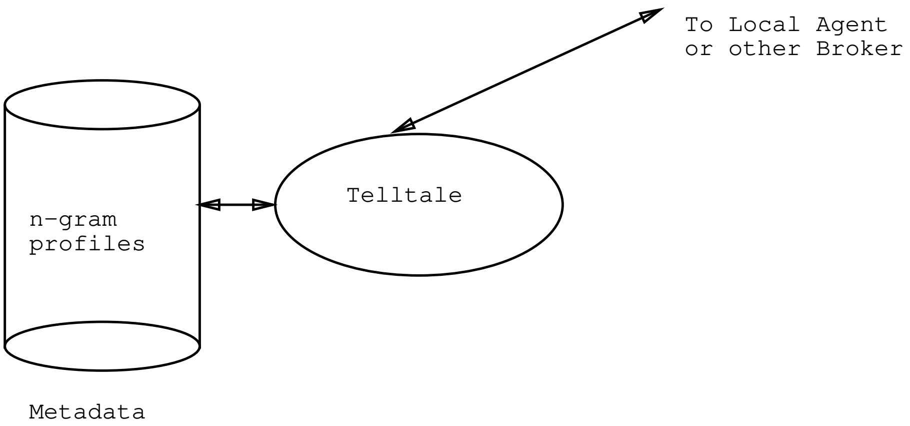

Figure 3: Broker

Telltale [Pea94, PN95] is an information retrieval system capable of building n-gram profiles. We use Telltale for managing n-gram profiles in both the local agent servers and in the brokers.

2.3 Brokers and Metadata

Returning to brokers and metadata, brokers are just like other agents, except that the data they manage is metadata supplied by various local agent servers, or other (sibling) brokers, or brokers above or below in a hierarchy. A hierarchy of brokers means arbitrarily large topologies are possible: since brokers are server agents in their own right, a broker of brokers can be imagined, and to however many levels are necessary

Brokers keep metadata, and use that metadata to route queries. The metadata collection serves as an implicit ontology, which could be augmented with an explicit ontology and a reasoning system, e.g. CYC [LGPPS90]. Note the range of sophistication in metadata possibilities.

In the mediated architecture described above, we envision that an agent calculates its n-gram profile from time to time, and transmits it to the designated broker(s). When a query arrives, brokers check the query against collected profiles and route query accordingly. See Figure 3.

3 Experiments with Telltale

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 4 for *The Wall Street Journal* corpus n-gram size vs. number of n-grams.

Number of n-grams as a Function of n-gram size	
n-gram size	number of 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 4: WSJ corpus size is approximately 40MB of text

Choice of n seems related to heterogeneity of the corpus – for a homogeneous corpus, larger n seems to give better retrieval effectiveness, measured as number of documents retrieved in response to a certain fixed set of queries. For a heterogeneous corpus, the number of documents retrieved in response to a fixed set of queries does not seem to depend on n.

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.

4 Scalability questions

Depending on the choice of n, an n-gram profile can be big, therefore the choice of n is important. Some investigation of the relationship between n and effectiveness of the metadata has been performed see 3. In addition, there is ongoing investigation into reducing the number of n-grams needed to effectively represent a corpus. Note that n-grams by their very nature contain much redundant information which is useful when working with degraded text (as from OCR input) but may not be necessary when text is relatively clean.

Because the "metadata question" can be answered with a probability, we can do probability-driven searches. In the worst-case, the metadata will direct an agent to query all other agents. While time and bandwidth consuming, this scenario is no worse than a single information agent managing the entire corpus, which is not possible due to size of the corpus vs. current size of storage media.

5 Conclusions and Future Work

In conclusion, we believe the system we have designed will scale to handle distributed information retrieval. The intelligent information system that we have described is partially implemented: we have KQML speaking agents, brokers, and agent name servers. We also have the text profiler and have used it for information retrieval on several individual corpora with good results. Now we will combine these two pieces to build a distributed intelligent information system with n-gram profiles as metadata to direct queries to appropriate agents.

References

[BDMS94] 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).

[BDHMSW95] 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.

[Bre94] Francis Bretherton. Reference Model for Metadata: A Strawman. IEEE Metadata Workshop, May 1994.

[Coh95] Jonathan Cohen. Highlights: Language- and Domain-Independent Automatic Indexing Terms for Abstracting. *JASIS*, April 1995.

[Dam95] Marc Damashek. Gauging Similarity with n-Grams: Language-Independent Categorization of Text. *Science*, Vol. 267, pp. 843-848, 10 February 1995.

[FLM95] Tim Finin, Yannis Labrou, and James Mayfield. KQML as an Agent Communication Language. to appear as a chapter in Jeff

Bradshaw (Ed.), *Software Agents* , MIT Press, Cambridge, to appear, 1995.

[LGPPS90] Douglas B. Lenat, Ramanathan V. Guha, Karen Pittman, Dexter Pratt, and Mary Sheperd. Cyc: Toward Programs with Common Sense. *Communications of the ACM Communications of the ACM*, Vol. 33, no. 8, pp. 30-49 (Aug 1990).

[Pea94] Claudia Pearce. A Dynamic Hypertext Environment Through *n*-gram Analysis. Ph. D. thesis, University of Maryland Baltimore County, 1994.

[PN95] Claudia Pearce and Charles Nicholas. TELLTALE: Experiments in a Dynamic Hypertext Environment for Degraded and Multilingual Data. To appear in JASIS, 1995.

[Sha48] Claude Shannon. A Mathematical Theory of Communication. *The Bell System Technical Journal*, Vol 27, pp. 379-423, 623-656, July, October 1948.