Resource Selection in CAFE: an Architecture for Network Information Retrieval

Grace Crowder *   Charles Nicholas

June 13, 1996

1 Introduction

The Intelligence Community is faced with the daunting task of extracting useful information from an ever-increasing supply of raw data. This raw data is captured in a variety of formats, but the focus of this work is on text: text that is written in a wide variety of natural languages, and which may be corrupt.

Computers have been used for text processing and information retrieval for many years, and there are a number of successful commercial systems. However, these systems are not up to the task that the Intelligence Community (IC) faces, for several reasons:

1. Scalability — the systems mentioned above are capable of dealing with corpora in the megabyte range, and small numbers of gigabytes in the case of Topic, but the IC needs systems that can scale up to text corpora in the terabyte range.

2. Heterogeneity — corpora tend to be partitioned on the basis of subject matter, source, estimated quality, date of collection, and so forth.

3. Multiple languages — the IC collects data in many natural languages, and many alphabets.

We believe a system in which subcorpora are managed by specialized agents may do better than any single, monolithic system could. To test this

*Computer Science and Electrical Engineering Department; University of Maryland Baltimore County; Baltimore, MD 21228; USA; {crowder,nicholas}@cs.umbc.edu

hypothesis, we are designing, building and evaluating the CAFE system. The acronym CAFE comes from "Cooperating Agents Find Everything." [1]

2 CAFE

2.1 Design Preconditions

The environment in which CAFE will be used imposes certain preconditions or assumptions on the design. Specifically,

1. Documents may be written in a number of languages, and may expressed in ASCII or UNICODE.

2. Typical documents range in size from 1k-100k. Some documents may be smaller or larger. Queries, in general, will be smaller than documents.

3. The corpus is at least several megabytes in size, and may grow to the terabyte range. Hence the number of documents in the corpus is expressed most conveniently in terms of thousands or millions.

4. The corpus is physically partitioned, and access to it is shared by a number of processors.

5. The corpus is dynamic, in the sense that new documents are entered on an essentially continuous basis.

6. The corpus may be very large, but it is *contained* in the sense of being under the control of a single organization or authority.

The last assumption has two implications. First, since the network of processors is under a central organizational control, it is possible (although not necessarily desirable) to install certain software on each network node. Second, we can assume a level of awareness about when and where new data comes into the system. These two factors make our environment different from the open Internet.

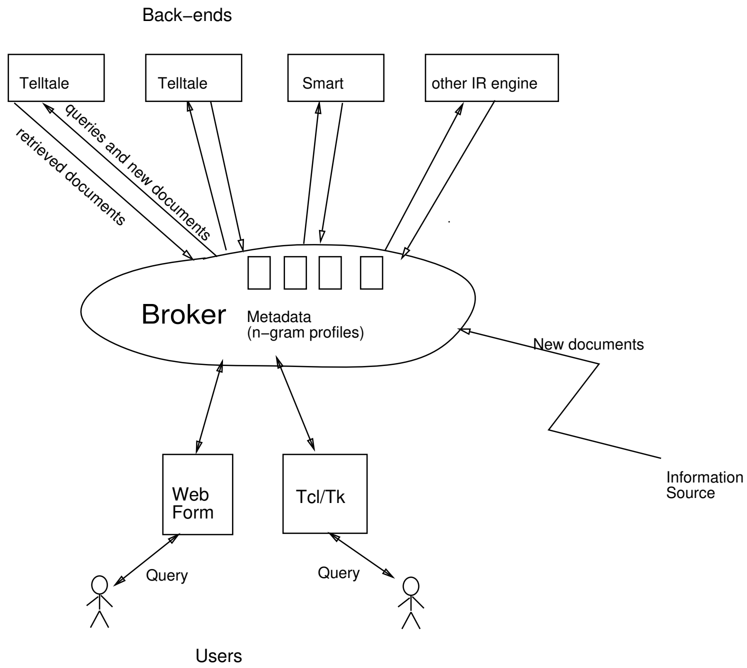

Figure 1: The CAFE system.

2.2 CAFE Architecture

CAFE, see Figure 1, is a large scale document information system consisting of agents managing local corpora using information retrieval engines as back-ends, brokers using metadata to direct information search, KQML (Knowledge Query and Manipulation Language) for agent communication, and a number of user interface agents.

We introduce the components of the architecture in turn:

- CAFE back-ends, see Figure 1, are information providers that manage local data and are capable of registering metadata with an assigned broker, responding to information requests from a broker, and updating corpora and corresponding metadata. To date, we have been thinking that an agent can be any information retrieval system that

[1] The CAFE project is part of a larger effort, referred to as the Massive Digital Data Systems (MDDS) program, being conducted under the auspices of the U.S. Department of Defense.

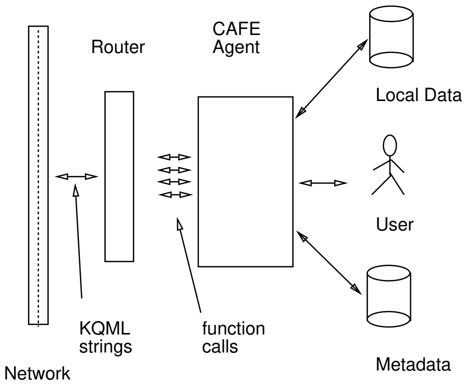

Figure 2: The generic CAFE agent has several components.

can accomplish these tasks. Agents will communicate with assigned brokers using KQML.

- KQML is an agent communication language based on speech–act theory [3]. In CAFE, a subset of KQML is used. Agents and brokers communicate to carry out basic information system needs. Agents register their metadata using the advertise performative, information requests are sent by brokers to agents using the ask performative. Results are returned to the broker using the KQML tell performative. The interesting use of KQML by CAFE is in using an information system as the agent locator in the broker.

- A CAFE *broker*, see Figure 1, is a KQML 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.

- Telltale is an information retrieval system based on the vector space model of information retrieval using n-grams as terms [4]. 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.

- 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.

  We are investigating the use of n-gram profiles as metadata. A frequency distribution of the n-grams in a document is called its n-gram profile. Pearce and Nicholas [4] 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 are effective, concise, automatically generated, abstractable, and interchangeable [1]. In addition, n-grams have been shown to be particularly effective when dealing with corrupt text data [4], and multilingual data [2].

3 Preliminary Results

To test different n-gram compression schemes, we'll take several scored corpora, and put them under the control of separate CAFE back-ends. The question to be addressed is, do the queries get handled correctly? We will look at two measures. First we will test if queries are routed to their associated corpora. Second, we will measure precision and recall for the whole of CAFE for each query.

We have implemented agents and brokers that route queries on a random basis. Such random routing represents a baseline against which other (more intelligent) techniques can be compared, and is a reasonable way to test the different components of the architecture. We are implementing a broker that uses Telltale and metadata to route queries. We are also implementing the n-gram compression schemes described in the next section.

We have experience with measuring precision and recall in Telltale in experiments involving varying n-gram size. We have also done some performance analysis of Telltale with respect to Smart [1].

4 Conclusions and Future Work

The innovative aspects of this work include the use of n-grams as metadata to aid in routing of queries and new documents. We are also interested in showing that this architecture gives reasonable performance, even for very large corpora.

We are looking at n-gram compression schemes to see how they perform. In particular, we are considering:

- using the top x n-grams,
- reducing the value of n (which reduces the number of unique n-grams),
- selecting n-grams on the basis of maximizing document coverage,
- sliding the n-gram window by other than 1 character.

For each of these approaches we are interested is determining whether compressed metadata, used by Telltale in the broker, is effective when used to route queries where data is scattered over numerous nodes

References

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

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

[3] James Mayfield, Yannis Labrou, and Tim Finin. Evaluating KQML as an Agent Communication Language. in Michael Wooldridge, Joerg P. Mueller, and Milind Tambe(Ed.), *Intelligent Agents II: Agent Theories, Architectures, and Languages*, Springer-Verlag Lecture Notes in AI - Volume 1037, 1996.

[4] 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.