

An architecture supporting the collection and monitoring of data openly contributed over the World Wide Web

David G. Stork
Ricoh California Research Center
2882 Sand Hill Road Suite 115
Menlo Park, CA 94025-7022
stork@OpenMind.org

Abstract

Open data collection over the World Wide Web — in which any web user can contribute to large databases of “informal” data — presents several challenges that require novel approaches in human interface design, algorithmic machine learning and collaborative infrastructure. Foremost among these challenges is the need to ensure data integrity and quality, by automatically or semi-automatically identifying unreliable or hostile contributors and rejecting their contributed data. Other, traditional requirements include security, scalability, and the need to support collaboration between separate but related data collection projects. Further, for the open software/data model, the system must permit users to freely browse, navigate and download the contributed data. This paper presents a collaborative architecture for a set of such data collection projects that addresses these challenges and requirements. Key components of this architecture have been implemented and tested as part of the Open Mind Initiative.

Keywords: Open Mind Initiative, open data collection, web infrastructure, open software development, data quality monitoring, anomalous data detection, learning with queries

1 Introduction: The need for large training datasets

Nearly all software projects in pattern classification and artificial intelligence (AI) — such as search engines and computer vision systems — require large sets of training data. For instance, state-of-the-art speech recognition systems are trained with hundreds or thousands of hours of speech sounds transcribed or “labeled” by knowledge engineers; leading optical character recognition systems are trained with pixel images of several million characters along with their transcriptions; one commercial effort at building a database of common sense information has required 500 person-years of effort so far, most of this in data entry [6]. In fact, there is theoretical and experimental evidence that given sufficiently large sets of training data a broad range of classifier methods yield similar high performance [4]; likewise, the vast majority of complex AI systems improve as their training sets are enlarged. In a Bayesian framework, we say that training with sufficiently large datasets “swamps the prior information” implicit in the choice of classifier model [2], and thus data becomes a dominant factor determining the final system

systems and a flightless classifier method. minor refinements in classifier algorithms and toward methods of collecting large reliable sets of accurately labeled data [5].

What are some of the sources of such vital data? Optical character recognition companies employ knowledge engineers whose sole task is to optically scan printed pages and then transcribe and confirm (“truth”) the identities of words or characters. Likewise, the Linguistic Data Consortium has dozens of knowledge engineers who transcribe recorded speech in a variety of languages on a wide variety of topics. Entering data by hand this way is often expensive and slow, however. An alternative approach, traditional data mining [3], is inadequate for many problem domains in part because data mining provides *unlabeled* data or because the data is simply not in an appropriate form. For instance the web lacks pixel images of handwritten characters and explicit common sense data and thus such information cannot be extracted by data mining. Moreover, accurately *labeled* data can be used in powerful *supervised learning* algorithms, while if the data is unlabeled only less-powerful *unsupervised learning* algorithms can be used [2]. For this reason, we naturally seek inexpensive methods for collecting labeled data.

As we shall see below, the internet can be used in a new way to gather needed labeled data: facilitating the collection of information contributed by humans. This paper discusses a novel large-scale collaborative framework for collecting data contributed over the World Wide Web — the Open Mind Initiative, as described in Section 2. The Initiative’s unique challenges and infrastructure are presented in Section 3. Sections 4 and 5 discuss the current implementation and future directions, respectively.

2 Open Data Collection

2.1 Trends in open software and infrastructure

There are several compelling lessons from collaborative software projects that have major implications for systems supporting the collection of data. Consider the open source software movement, in which many programmers contribute software that is peer-reviewed and incorporated into large programs, such as the *Linux* operating system. Two specific trends must be noted. The first is that the average number of collaborators per project has increased over the past quarter century. For instance, in the late 1970s, most open collaborative software projects such as *emacs* involved several hundred programmers at most, while by the 1990s projects such as *Linux* involve over 100,000 software engineers. The second trend is that the average technical skill demanded of contributors has decreased over that same period. The programmers who contributed to *gcc* in the 1980s were experts in machine-level programming; the contributors to *Linux* know about file formats and device drivers; the contributors to the *Newwho* collaborative open web directory need little if any technical background beyond an acquaintance with *HTML*.

These trends, and the emergence of the World Wide Web, suggest that collaborative efforts can be extended to an extremely *large* pool of contributors (potentially anyone on the web), whose technical expertise can be *low* (merely the ability to point and click). This is the approach we explore below.

2.2 The Open Mind Initiative

The central goal of the Open Mind Initiative is to support non-expert web users contributing “informal” data needed for artificial

intelligence and pattern recognition projects. The Initiative thus extends the trends in open source software development to larger and larger groups of collaborators, allowing lower and lower levels of technical expertise. Moreover, the Initiative broadens the output of collaborative projects: while traditional open-source projects release software, the Initiative releases both software and data [8]. Our goals also differ markedly from `gnutella.org` and `napster.com`, which seek only to share existing files, not *learn* from or generate new ones.

A prototypical open data collection project in the Initiative is illustrated in skeleton form in Fig. 1. The project site contains a large database of isolated handwritten characters, scanned from documents, but whose character identities are not known. Individual characters from this database are presented on standard web browsers of contributors who then identify or “label” the pattern by clicking buttons on a simple interface. These labelings are automatically sent to the project site, where they are collected and used to train software that classifies handwritten digits.

Some data acquisition projects in the Initiative could employ novel human-machine interfaces based on games. For instance, imagine an Open Mind Initiative chatbot project in which data is collected while contributors play a modified version of *Dungeons and Dragons*. In this new game, players read short texts — which discuss potions to drink, swords to brandish, rooms to enter, tasks to accomplish — generated by automated text generation programs. As part of the game, players must indicate how “natural” these texts are. This valuable feedback, collected at the project site, provides information for adjusting the parameters in the text generation programs, thereby yielding more natural generated text. In such game-based projects, contributors download the game software (presumably written in *Java*) from the project site.

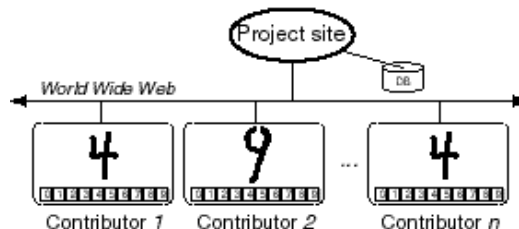


Figure 1: This simplified, skeleton architecture shows the general approach in an open data collection project on isolated handwritten digit recognition. The unlabeled pixel images are presented on the browsers of contributors, who indicate their judged memberships by means of a button interface. Occasionally, the same pattern is presented to two or more independently selected contributors, to see if they agree; in this way, the reliability of contributors is monitored semi-automatically.

The data captured on the contributor’s machine is stored locally and sent to the project site at the end of a game session.

There are a number of incentives for people to contribute to Open Mind Initiative projects. Contributors seek benefit from the software (as in a text-to-speech generator); they enjoy game interfaces (as in online versions of *Dungeons and Dragons*); they seek public recognition for their contributions (as in SETI@home); they are interested in furthering the scientific goals of the project (as do amateur ornithologists through annual bird counts for the Audubon Society); they seek financial incentives such as lotteries, discounts, e-coupons or frequent-flier awards provided by third-party corporations [7].

The Open Mind Initiative differs from the Free Software Foundation and traditional open-source development in a number of ways. First, while open-source development relies on a hacker culture (e.g., roughly 10^5 program-

mers contributing to *Linux*), the Open Mind Initiative is instead based on a non-expert web user and business culture (e.g., 10^9 web users). While most of the work in open-source projects is directly on the final software to be released (e.g., source code), in the Initiative most of the effort is directed toward the tools, infrastructure and data gathering. Final decisions in open source are arbitrated by an expert or core group; in the Initiative contributed data is accepted or rejected automatically by software that is sensitive to anomalies or outliers [1]. In some cases, data can be rejected semi-automatically, for instance by having data checked by two or more independently chosen contributors. Such “self-policing” not only helps to eliminate questionable or faulty *data*, it also helps to identify unreliable *contributors*, whose subsequent contributions can be monitored more closely or blocked altogether. It must be emphasized that the Open Mind Initiative’s approach also differs significantly from traditional data mining. In particular, in data mining a fixed amount of unlabeled information is extracted from an existing database (such as the web), whereas in the Initiative a possibly boundless amount of labeled data is *contributed*.

The Open Mind Initiative has the potential to be one of the largest collaborative projects of any sort, with anyone on the web a contributor (www.OpenMind.org).

3 Infrastructure for open data collection

The goals listed above place a number of constraints on the infrastructure of the Open Mind Initiative. Figure 2 shows a schematic of the Initiative’s infrastructure, which supports computation, collaboration, and a range of database functions. Central is the home

web site which contains background information and documentation for potential contributors, the mailing list and archive, as well as mirror databases of the data collected at individual projects. In the open source/data model, such mirroring gives added confidence to contributors that the data is available to all and cannot be controlled by a single person or group. Automatic ftp transfers at regular intervals update the usage logs and these mirror databases at the home site. Each project site has software for monitoring data quality and identifying unreliable or hostile contributors. The email addresses of such unreliable contributors are forwarded to the Initiative home site and are available to other projects. Since the previously contributed data is indexed by the identity of its contributor, data from discredited sources can be removed at any time.

While in most of the Initiative’s projects contributors provide data through standard web browsers, in other projects contributors will require a more sophisticated human interface. For instance, in projects using a game interface, contributors will download the presentation and local caching software resident from the project site, and install it on their local machine. Data is collected while the contributor plays the game and is sent to the project home site at the end of a game session. This organization reduces the direct web connection time for contributors, which might be of concern for contributors paying for dialup access to the web.

In an open source software/data framework, the infrastructure and architecture also support external experts, for instance the developer of handwriting recognition systems who seeks the Open Mind Initiative’s labeled character database. Such an external expert would have to digitally sign an open source/data license, and would then be free to download the data from either the project site in question or from the mirror database at the home site.

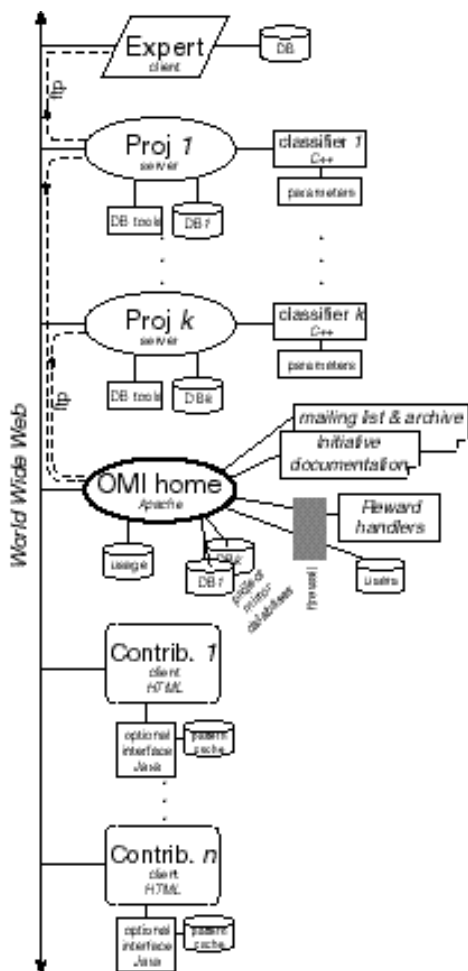


Figure 2: System architecture of the Open Mind Initiative, here showing the home site, k project websites, n desktop machines of non-expert contributors, and one external expert.

The Initiative also profits from the contributions of programmers of infrastructure, database and security software, much as in the *Linux* developer community. So far such contributions represent a small proportion of the Initiative’s total effort. The current architecture has no special provisions for facilitating contributions of software, which are handled by each component project on an

ad-hoc basis. As contributed software becomes a greater consideration in projects, well-established open-source methods will be used for sharing software across the entire Initiative [9].

While newcomers to the Initiative generally go to the home page first and then follow a link to a project of particular interest, in later sessions they can go directly to the chosen project. As such, the scalable distributed architecture in Fig. 2 thus reduces total network traffic and increases the overall speed of communication.

This infrastructure and software provide a model for other data collection projects — on the web or on local area networks.

4 Current status

The Open Mind Initiative has four projects in progress: handwriting recognition, speech recognition and a small, demonstration AI project, *Animals*. These have been tested on intranets and are being debugged and load tested for full web deployment. The fourth project site, Open Mind common sense, is open and accepting contributed data over the web. As of April 2001, it has collected 400,000 common sense facts from 6500 separate contributors through a range of “activities,” such as DESCRIBE A PICTURE and RELATE TWO WORDS. To date, data monitoring in this project has been semi-automatic whereby contributors “self-police” the contributions of each other.

Much of the infrastructure shown in Fig. 2 is in place and has been tested, such as the documentation, mailing list and archive at the home site, links to individual projects and databases at project sites. While each project leader has software and tools for navigating their own databases, these tools have not yet been refined for external users.

5 Future work

There remains much work to be done on the Open Mind Initiative itself and its infrastructure. Indeed, like most open-source projects, the Initiative is an ongoing and evolving endeavor. Our next major milestone is to load test the Open Mind handwriting project and transfer it to the web. The subsequent step is to develop software for smooth navigation and integration of project databases.

We are soliciting additional projects that would profit from the Open Mind Initiative approach, for instance projects for collecting the parts of speech of sentences extracted from the web (so-called “bracketing”), game-based interfaces, and so on. A particularly intriguing project would be to collect digital images of animals, to be used to train sophisticated computer object recognition algorithms. Thus a contributor would upload JPEG images of pets or forward ones already on the web. We are developing machine learning algorithms for automatically monitoring the quality of contributed data and we believe the current architecture will support such algorithms.

The general infrastructure of the Open Mind Initiative is flexible and scalable; it should support a range of new projects and requirements as they are introduced.

Acknowledgements

I thank Push Singh (MIT Media Lab) and Steve Savitzky (Ricoh) for useful discussions.

References

- [1] Kari Alho and Reijo Sulonen. Supporting virtual software projects on the web. In *Seventh IEEE International Workshop on Enabling Technologies: Infrastructure for Collaborative Enterprises (WET ICE)*, pages 10–14, Stanford, CA, USA, 1998. IEEE Computer Society Press.
- [2] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. Wiley, New York, NY, second edition, 2001.
- [3] Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramaswamy Uthrusamy, editors. *Advances in knowledge discovery and data mining*. MIT Press, Cambridge, MA, 1996.
- [4] Tin Kam Ho and Henry S. Baird. Large-scale simulation studies in pattern recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-19(10):1067–1079, 1997.
- [5] Chuck Lam. *Open Data Acquisition: Theory and Experiments*. PhD thesis, Stanford University, Stanford, CA, 2002. in preparation.
- [6] Doug B. Lenat. CYC: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38, 1995.
- [7] David G. Stork. The Open Mind Initiative. *IEEE Intelligent Systems & their applications*, 14(3):19–20, 1999.
- [8] David G. Stork. Open data collection for training intelligent software in the Open Mind Initiative. In *Proceedings of the Engineering Intelligent Systems Conference (EIS2000)*, Paisley, Scotland, 2000.
- [9] Davor Čubranić and Kellogg S. Booth. Coordinating open-source software development. In *Eighth IEEE International Workshop on Enabling Technologies: Infrastructure for Collaborative Enterprises (WET ICE)*, pages 61–65, Stanford, CA, USA, 1999. IEEE Computer Society Press.