

KD-D Statement of Work:

Large-Scale Visual Interaction with Multi-Lingual Multi-Source Textual Data

Gregory B. Newby, University of North Carolina at Chapel Hill

Vision: To monitor and retrieve from hundreds of changing data sources using a variety of data representation schemes. Analysts and other information seekers will set up profiles for scanning the incoming data, and utilize powerful visualization tools for assessing trends, anomalies and new events. The vision of this proposal is a highly interactive visual environment for experiencing a changing information landscape.

Overview: Techniques for visualizing, mining and retrieving textual data have seldom been applied to dynamic streams of data. From a first year emphasizing three languages with data from news media sites, this research project strives to distill a very large number of discreet data sources to a single filtering and display system.

Background: Information retrieval is not a solved problem. Despite the success of modern Web search engines and database systems, information seekers have only limited ability to express their unique information needs. While cross-language information retrieval (CLIR) has promise, research systems for cross-language retrieval suffer from not being tested in operational environments with ongoing changeable streams of data.

This research will integrate best practices from CLIR, text filtering and retrieval. It will implement advanced visual interfaces for understanding the flow of textual data, and harnessing that flow. Contemporary news articles and other data in English, Arabic and Mandarin Chinese will be used to build a parallel corpus (collection). In this corpus, pairs of articles about the same topic, but in different languages, will be identified. Latent semantic indexing (LSI) and variations will be used to enable the matching of a query in any of the three languages to documents in any of the three languages.

There are significant scientific and practical challenges to this work, but the work is firmly grounded in established techniques. Challenges include:

1. Practical issues in gathering data for a parallel corpus, including data quality, granularity, named entities, and topic detection. Automatic selection of parallel documents and augmentation of corpus building by the use of machine translation are desirable.
2. Visualization tools for textual retrieval and filtering, and assessment of their effectiveness. Many 2D and 3D tools have been developed, but almost none have been implemented in environments with “live” data. Dominant models will be implemented and evaluated, including point-cloud spaces (scatter plots) based on LSI, Kohonen self-organizing maps, and others (note that the [PNNL](#) has outstanding prototypes of similar tools, but generally without “live” data).

3. Scaling for LSI and related techniques. The computational complexity of large parallel corpora can be daunting. Incremental methods for adding new documents and recomputing the information space will be assessed, and existing high-performance algorithms for LSI, eigensystems, multivariate regression etc. will be adapted and compared.
4. Evaluation of the utility of the major aspects of this work, alone and combined, is critical. Different information seekers and different information needs are envisioned, and their fit with the variety of systems approaches must be assessed for real-world benefits to be realized.

Data (year 1): Professional media Web sites from newspapers, television and Web-only broadcasters will be monitored. Human experts will collect instances of the same story in different languages to create a parallel corpus to facilitate cross-language information retrieval. The first year will use English, Arabic and Mandarin Chinese (or other languages if requested by funders). By training and evaluating automated agents, the cross-language collection building will be partially or entirely automated by the end of this year. Only publicly available (free or by subscription) data sources will be utilized.

Data (years 2-3): Additional data sources will be added, especially those with more rapidly changing content. Online chats and mailing lists or newsgroups will supplement media Web pages. Streams of data from text-to-speech conversion of radio programs and closed-caption data from television broadcasts will also be added. More languages will be added. Only publicly available (free or by subscription) data sources will be utilized, unless other data sources are made available by the KD-D program sponsors.

Milestones (year 1): The PI's techniques for large-scale information retrieval will be applied to the tri-lingual corpus. Visualization techniques based on latent semantic indexing (LSI), self-organizing maps (SOM), statistical trends analysis and data abstraction will be integrated in an information space visualization tool. Deliverables will include:

1. A tri-lingual parallel corpus of thousands of items, covering contemporary events and personalities.
2. Profile-based filtering methods for identifying new items of interest from the daily stream of news, including query by example and keyword search.
3. Visual tools for examining the relations among data items, graphical query specification, and trends analysis.

Milestones (years 2-3): Further visualization techniques, more data, user-based evaluation, and tuning for particular information needs or information seekers. Details on deliverables will be further specified as part of the year 1 activities, and will include:

1. Integration of additional languages (Japanese, German, Thai or others, based on availability of data sources, a user community for evaluation, and interest. Note that additional languages should pose no problem for the parallel corpus and LSI

- technique, as it is not necessary to have a document in *all* languages, only two or more.
2. Incorporation of more dynamic data sources, with poorer segmentation. Existing speech-to-text techniques for radio news broadcasts (via short wave and the Internet) and television (using speech-to-text or closed-caption data) will produce a dynamic data stream. Other sources, including non-text sources, will be tested.
 3. A full visualization suite for textual data will be developed. Many different methods for textual data visualization are available, with different metaphors (data landscapes, point-clouds, data histograms, etc.). A toolkit for textual data visualization will be implemented and evaluated.
 4. User-based evaluation of the information filtering and retrieval systems, along with the visualization toolkit. Real-world information need situations or scenarios will be sought via public testing or (if available) from partner agencies. Information seeker profiles will form a basis for tuning the visual retrieval environment for particular query types.

Researcher team credentials: The PI has developed, and is continuing to develop, high-performance tools for large-scale text retrieval ([IRTools](#)). The purpose of IRTools is to enable information scientists to perform evaluation studies that will inform practice. Unfortunately, many retrieval techniques that work in the laboratory have never seen use in practical systems, due to the lack of stable software and resolution of scalability issues. The PI has also worked extensively with visualization methods for text retrieval, especially point-cloud systems derived from LSI-type techniques. Usability studies of these systems indicate that 2D and 3D interfaces are functional, but require additional training and do not necessarily replace traditional text-based interfaces (e.g., typed queries and ranked list of documents). The PI intends to work with colleagues who are leaders in information visualization (Alan Hudson, chair of the Web3D group and VRML designer), user-based evaluation and agile views (Gary Marchionini at UNC) and optimal computational methods for LSI and related techniques (Michael Berry at UTK).

Budget: Full details are provided in a spreadsheet included with this proposal. The total cost for one year is approximately \$400,000.

Supplemental documents: For related literature, please see <http://ils.unc.edu/gbnewby/kdd/kdd-lit.html>. For budget details, a spreadsheet has been provided, and is available as <http://ils.unc.edu/gbnewby/kdd/kdd-budget.xls>.

Starting date: A May 1 start date is proposed. This will allow time for hiring and acquisition of materials, and corresponds with the start of the summer session at UNC.