New contextual data mining technologies

Michael W. McGreevy

Four new contextual data mining technologies have been developed and tested, and they are in the process of being patented and commercialized by Ames. The methods were developed as part of NASA's effort to dramatically reduce the potential for commercial aviation accidents by modeling and monitoring safety-related incidents in the National Airspace System. Specifically, this work provides a new set of information technologies that automatically interpret the contextual structure of incident and accident reports to develop detailed, computable models. These models represent the domain of commercial aviation operations, the situations described, and the concerns of the incident and accident reporters. These new methods provide greatly improved information retrieval from large databases of unstructured text.

The four new data mining methods together are called QUORUM/Perilog, and are comprised of keyword-in-context search, phrase search, phrase generation, and phrase discovery. These methods build upon a common core of contextual analysis, modeling, and relevance-ranking of text. QUORUM/Perilog keyword-in-context search retrieves text, such as incident narratives, that contain one or more user-specified keywords in typical or selected contexts, and ranks the narratives on their relevance to the keywords in context. It displays those narratives with their relevant sections highlighted, and also displays the criteria used to determine relevance. QUORUM/Perilog phrase search retrieves narratives that contain one or more user-specified phrases, even hundreds of phrases, and ranks the narratives on their relevance to the phrases. It displays the narratives with the phrases highlighted, and also shows near-matches to the query phrases. QUORUM/Perilog phrase generation creates a list of phrases, from the database of text, that contain a user-specified word or phrase. QUORUM/Perilog phrase discovery finds phrases that are related to topics of interest. For example, a query on the topic of "fatigue" produces results including: "rest period", "reduced rest", "crew rest", "continuous duty", "crew scheduling", "duty period", "rest periods", "reserve or standby", and many others. Phrase discovery is useful for gaining a better understanding of the topics contained in a database. In addition, phrase generation and phrase discovery are particularly useful for finding query phrases for input to QUORUM/Perilog phrase search.

The new data mining methods have been successfully tested on the tens of thousands of narrative incident reports in the database of the Aviation Safety Reporting System (ASRS). The ASRS is a national clearinghouse of aviation safety incident reports supported by the FAA and managed by NASA. The new technologies completed this year will support system-wide ASRS analyses for government, industry, and academic researchers.

The new data mining methods are applicable to a wide variety of text, and several analyses have been done as demonstrations. For example, one analysis involved reports of safety-related incidents that occurred during ground maintenance of the Space Shuttle. In addition, a sample of incidents from the electric power generation industry was analyzed.

Commercialization of the four new data mining methods began with the submission of four formal invention disclosures. The Ames Commercial Technology Office (CTO) reviewed the disclosures, and also hired contractors to conduct a search for prior art and an analysis of the commercial potential of the new technologies. Encouraged by the results, the CTO established a contract with an outside law firm to write four patent applications. The four patent applications are based on nine "inventive concepts", each of which could be an independent patent if the intent were to maximize the number of patents, as is often done in start-up companies. The writing of the patent applications was still in progress at the end of Fiscal Year 2000. After approval by Ames and Headquarters, completed applications are submitted to the United States Patent Office. Meanwhile, the CTO is developing a commercial licensing strategy for the new technologies.

The new methods have been documented in the paper, "Searching the ASRS database using QUORUM keyword search, phrase search, phrase generation, and phrase discovery". The paper has been approved for publication pending submission of the patent applications.