The bioPrompt-box: an ontology-based clustering tool for searching in biological databases

Corsi C, Ferragina P, Marangoni R

Department of Informatics, University of Pisa, Pisa.

Motivation

High-throughput molecular biology provides new data at an incredible rate; the increase in the size of biological databanks is enormous and very rapid. This scenario generates severe problems not only at indexing time, where suitable algorithmic technologies for data indexing and retrieval are required, but also at query time, since a user query may produce such a large set of results that their browsing becomes unaffordable. This problem is well known to theWebIR community and, in fact, third-generation search engines are currently providing new tools to better and better satisfy the "user needs" behind their queries: personalization (e.g. Eurekster), user behavior profiling (e.g. Google, Yahoo), query term suggestion (e.g. AskJeeves), are just a few of them. Web-results clustering is another approach pioneered by NorthernLight (1996) and then recently made famous by Vivisimo.com. The basic idea is that the results returned by a (meta-)search engine are clustered into folders which are labeled with intelligible sentences capturing the "theme" of the results contained into them. The folders are further organized in a hierarchy, whose internal nodes are labeled with sentences too. Users can therefore navigate through the labeled folders driven by the "need" behind their query. In this way (lazy) users are not limited to look at first ten results (more than 85% of web users do it), but they can immediately acquire several points of view on a larger pool of them. Consequently, they can either narrow their search by clicking on some folders, or they can acquire new knowledge by looking at the folder labels, and possibly refine their query. These nice features have driven some authors to say that "clustering technology is the future of search engines"

Methods

In this work we present a tool, TheBioPrompt-box, that follows the labeled hierarchical clustering approach in the context of biological databanks (specifically Uniprot, for now), thus exploring the applicability of this technology to the specialties of the biological data, with the ultimate goal of making the exploration task of biologists easier, more effective and more efficient. It goes without saying that, while Web pages are heterogeneous and uncontrolled so that their clustering must operate on-the-flight over the text excerpts returned by the queried search engines (cfr. Vivisimo.com); in the biological context, data are metatagged and come from multiple (humanly controlled) sources. As a result, the clustering and labeling task is, from one side, simplified by the availability of these metadata but, from the other side, it is made more challenging because it needs new methods for combining these sources into more effective labeled clusters. This is exactly the scenario we explore with TheBioPrompt-box which, at the best of our knowledge, is the first system in the literature adopting this kind of approach on biological data. In fact, ClusterMed is a similar project, but it offers a clustering only on the bibliographic database PubMed.

Results

At this stage, TheBioPrompt-box defines a meta-document as a sequence (protein or gene) plus the meta-data associated with that sequence in UNIPROT. So, a (meta)document is a protein sequence with the information about the organism, the comments of the researchers, the references to Gene Ontology, the references to articles or publications related to this sequence (possibly referring to Pubmed), or the taxonomy lineage of the organism. This means that the (meta) documents indexed by TheBioPrompt-box are composed by a list of fields containing either references to ontologies, or to external databanks, or they are plain text like researcher comments and (title of) articles. TheBioPrompt-box indexes this (meta) documents collection using Apache Lucene, a highperformance, full-featured and open-source text search engine library written entirely in Java, in connection with Apache Commons Digester to manage XML format data. At query time,

TheBioPrompt-box offers to the user some useful tools to customize the search and the clustering process. The more relevant query types currently supported are: Free Text over all fields of the (meta) document; Sequence-based which exploits Blast; and Keywords-based which searches over the keyword field of each indexed (meta) document. The search process is intended to select a set of (meta) documents over which the clustering algorithm is then executed. Currently, the user may cluster the (meta-)documents according to three different rules: the terms of Gene Ontology the (meta) documents cite, their lineage, or the organism information they share.

Availability: http://brie.di.unipi.it:8080/BioPrompt-box/

Contact email: marangon@di.unipi.it