

# Statistics Topics

Fei Yu

Advisor: Jim Pitman

May 5, 2010

## 1 Overview

In this project, we created a richly-linked web portal, called *Statistics Topics Index* (STI), for displaying statistics topics. In addition to parent topics, child topics, and similar topics, a topic may also be linked to practitioners, courses, and textbooks. All the information pertaining to a topic will be aggregated when STI presents the topic to visitors.

STI is the first step in our attempt to better understand the relationships between the network of statistics topics and its topologies.

## 2 *Statistics Topics Index*

*Statistics Topics Index* (STI) is a richly-linked web portal for displaying statistics topics. STI consists of two major components: a network of statistics topics and its topologies. Currently, the topologies include tables of contents of statistics textbooks, lecture summaries from statistics courses, and statistics practitioners' research interests.

In STI's network of topics, each topic interrelates with the others based on one of the following four types of relationships: "parent-child", "child-parent", similar, and unrelated. In addition, each topic may be linked to tables of contents, lecture summaries, or research interests. Furthermore, the arrangement of topics are subject to different topologies.

Visitors of STI may browse the collection of statistics topics directly, or view all the topics affiliated with a certain practitioner, course, or textbook. Besides offering a basic navigation interface for each topology, STI emphasizes on providing the visitors with richly-linked displays of statistics topics. For example, while browsing through a topic's webpage, one finds not only the information about the topic, but also references to its parent topics, child topics, similar topics, and all other sources, such as tables of contents, in which the topic is mentioned.

While STI is self-contained with its own network of topics and the topologies, we expand the wealth of information in STI by incorporating pertinent information from the Internet. Right now, STI redirects visitors to search engines and encyclopedias with scripted searches. In the future, however, we plan to filter and parse search results from the Internet, and display them directly on STI.

## **2.1 Data in STI**

STI currently contains over 5500 statistics topics. They come mainly from UC Berkeley's school catalog, International Statistical Institute's glossary of statistical terms, Mathworld, Mathematics Subject Classification, textbooks' tables of contents, and Professor Jim Pitman's lecture notes.

## 2.2 Data Schema and Data Storage

Part of the challenge of building such a richly-linked web application is to find a data schema suitable for fast information aggregation, that is, aggregating all the sources in which the topic is referenced, and easy modification, such as adding and removing topologies. In the project, we developed STI on two different data schemas based on two different databases—MySQL, a relational database, and CouchDB, a document-oriented database. The schemas are referenced in Section 4.

## 2.3 MySQL vs. CouchDB

STI was first built on MySQL, a relational database, and was later built on CouchDB, a document-oriented database. Although MySQL provides a robust data structure for STI, as all the relations between data are well-specified as a requirement by relational databases, CouchDB offers the flexibility of introducing new data and changing current the data structure without having to modify existing data. As a result, it's more convenient to use CouchDB during project development. Furthermore, because of CouchDB's leniency on data structure, it is expected that the CouchDB-based STI can be easily adapted for similar web applications.

## 2.4 Extensibility

Judging from STI's performance with the current set-up and data schema, we expect STI to be a highly portable web application—a web application that can not only be deployed in different machines, but can also be adopted for disciplines other than statistics. Furthermore, we expect STI to be scalable, as it handles the current data with no noticeable latency.

### 3 Outlooks

STI may prove attractive to students and researches who appreciate an information aggregation portal, but the main purpose of creating STI is to provide the infrastructure for analyzing the relationships between the network of statistics topics and its topologies. Some investigations of statistics topics, such as inferring the relationships between statistics topics from tables of contents, have been done using STI.

The future of the Semantic Web, which includes RDF, RDFS, OWL, and SKOS, appears promising. And the Semantic Web's highly structured data fit STI's data requirements nicely. Therefore, we hope to incorporate the Semantic Web into STI, providing importing and exporting of Semantic Web data, and possibly using it as a data storage and dissemination solution.

Future developments of STI will continue to focus on Statistics topics. As STI places no restriction on the content of its data, we expect similar fields such as Mathematics to adopt STI's relationship management and data visualization tools.

### 4 Code

1. Project wiki
2. Data: CouchDB, MySQL
3. Data Schemas: CouchDB, MySQL
4. Source code: CouchDB, MySQL

## 5 Contributions

In addition to guidance from Professor Jim Pitman, STI represents the collective effort of Jeff Regier [[jeff@stat.berkeley.edu](mailto:jeff@stat.berkeley.edu)], a graduate student in the Department of Statistics, UC Berkeley, and myself. Jeff Regier created the original topic data model in STI, which includes inter-topic relationships and scripted searches, collected the topics and authors data, and set up MySQL and the server to support STI. I extended the data model in STI to include various topologies such tables of content and lecture notes, and adapted STI to a CouchDB-backed system.