Interesting Tasks for Restaurant Discussion Bulletin Boards

Jason D. M. Rennie

July 13, 2004

Informal communication, such as e-mail, bulletin board posts and mailing lists, is unique in that it represents an information resource at least as large as the Web, yet it is little used. Some mailing lists and bulletin boards are available on the Web, but traditional bag-of-words and link-based indexing method do not do justice to the information they contain. A major difficulty with informal communication is context. Web pages tend to be more stand-alone than e-mails, and yet it is Web pages that have easily-harvestable context information: link structure. E-mail has plenty of context, but it is in the form of natual langauge; extracting context information requires (some degree of) natual language understanding. The context problem is key in solving other important information problems involving informal communication, such as search and information extraction. Link structure was instrumental in making Web search useful—it provided essential information about context. Web pages tend to mention their topic, but do not generally follow the bag-of-words model (number of mentions in proportion to amount of discussion on the topic). An extensive discussion on Armadillos may only include a few mentions of "Armadillo." Once context is established, language conventions dictate that there is no need to repeat the context over and over again. The text in and around Web links provide general topic information. So, searching on the text in links pointing to a Web page can be more useful than searching the actual text of the Web page. Modern search engines use a mix of Web page text and link text to determine the relevance to a search. Modern-day mailing list search is similar to the way search was in the early days of the Web: link structure was ignored. Since link structure is generally not available in e-mail and Web bulletin boards, we must extract context from the language itself.

Being able to determine context is a necessary requirement for a number of informal communication tasks. It aids search in the same way that in-link text helps Web search; it provides what the actual text doesn't: topic keywords. Search conducted over a combination of context keywords and document text will be more successful than over the document text alone. There are many other tasks for which context provides useful, if not essential information. The tasks vary by application, so we consider the specific problem of extracting information from a restaurant review bulletin board. A task closely related to the task of determing context is the identification of restaurant names. Most discussions center around one or a set of restaurants. Identifying restaurant names is similar to identifying named entities in other applications. Context plays a key role in a related task: resolving references to restaurants. Once a restaurant name is mentioned once, it is often later referred to using pronouns and/or abbreviations. The most difficult resolution cases involve identifying the context (the restaurant being discussed). A related task is identifying new restaurants (or new named entities). This highlights the need for correctly resolving shorted, abbreviated and misspelled forms. Other domains have similar issues. For example, the presidents Bush and the Iraq Wars. A task that is somewhat special to restaurants is the detection of major changes in quality, ownership and/or management. When a restaurant gets a new chef and/or management, it may retain its old name, but be "new" for all practical purposes. It is important to be able to distinguish between the different versions of the restaurant. It may be useful to be able to make this distinction for various types of organization, such as businesses, countries, hotels, etc. The recent change of power in Iraq provides a good example

We have established context as an important building-block for extracting information from informal communication. A second building-block is the categorization of statements. It is useful to be able to separate questions from statements and factual statements from opinions. It is also useful to breakdown opinions and statements based on their subject matter, e.g. food vs. service vs. ambiance opinions of a restaurant. Factual statements can be used to categorize the restaurant as to the type of food and the general class of service and/or formality. Finally, we can compile summary opinions of a restaurant by grouping individual opinions by subject matter and/or sentiment. For example, to get an overall rating score for a restaurant, we can determine, for each user, whether his/her opinions are positive or negative; averaging this sentiment across users gives an overall score. By further categorizing each opinion, we can obtain summary opinions by subject matter. Establishing rank or numerical scores for each user's view of each restaurant allows us to leverage collaborative filtering: we can recommend restaurants the user might enjoy based on how his/her reviews line up with other users.