Name of Scholar: Ahmad Kamal

Name of Supervisor: Dr. Muhammad Abulaish

Department: Computer Science, Jamia Millia Islamia, New Delhi

Title of Thesis: Mining Web Opinion Sources Using Machine Learning Techniques

---

Abstract

Due to increasing popularity and cheap access of internet, the World Wide Web (WWW) is emerging as a new medium to share individual experiences or opinions. Web opinion sources are rapidly emerging in the form of merchant sites, forums, discussion groups, and blogs, attracting individual users to participate more actively for sharing their experiences or opinions. Experiences of existing users are very useful for new ones to help them to choose right products, and for product manufacturers to know the strengths and weaknesses of their products from users' perspectives. As the number of reviews that a product receives may grow rapidly and many times the reviews may also be quite lengthy, it is hard for the customers and manufacturers to analyze them through manual reading to make an informed decision.

Since last decade, the application and usage of opinion mining, especially for business intelligence, have fascinated many research attentions around the globe. Various research efforts attempted to mine opinions from customer reviews at different levels of granularity, including word-, sentence-, and document-level. Document-level sentiment analysis classifies a review document as positive, negative, or neutral and fails to provide insight about users' sentiment on individual features of a product or service. Therefore, it seems to be a great help for both customers and manufacturers, if the reviews could be processed at a finer-grained level and presented in a summarized form through some visual means, highlighting individual features of a product and users sentiment expressed over them.

In this thesis, starting with the design of a rule-based system to identify candidate information components, we have proposed various machine learning techniques for subjectivity/objectivity analysis, feasible feature-opinion pairs identification, and sentiment classification. We have also proposed the design of a novel feature-level review summarization scheme to visualize mined features, opinions and their polarity values in a comprehendible way.

The proposed rule-based system applies linguistic and semantic analyses on review documents to identify feature-opinion pairs based on Parts-Of-Speech (POS) information and dependency relationship generated by a statistical parser. A rich set of discriminative features based on statistical and linguistic characteristics of review documents is identified for learning classification systems to identify subjective sentences from review documents. Filtering out objective sentences, which generally contains factual information, drastically enhances the accuracy of opinion mining system. It has been observed that product features mentioned in a sentence are referenced in succeeding sentences using anaphoric pronouns. In order to identify the association of features with correct opinions presents in another sentences, a backtracking-based anaphora resolution approach is presented for correct binding of feature-opinion pairs.

To eliminate noisy feature-opinion pairs, a bipartite graph is generated considering feature-opinion pairs as hubs and review documents as authorities, and HITS algorithm is applied to calculate reliability scores for feature-opinion pairs; a higher value of which reflects a tight integrity between the components of a pair. Besides rule-based approach, an alternative *n*-gram based supervised machine learning approach is proposed to minimize the use of dependency relationships generated by statistical parser for extraction of information components. Based on a rich set of discriminative features, a binary classification system is learned for valid feature-opinion pair identification.

We have also proposed a rich set of statistical features including *Pointwise Mutual Information*, *Mutual Information*, *Chi-Square*, and *Log Likelihood Ratio*, in addition to some linguistic features, including *negation*, *tf-idf*, and *modifier* to model a word-level supervised machine learning sentiment classification system, which determines the polarity of opinionated words as *positive*, *negative*, or *neutral*.

Finally, we have proposed the design of an opinion summarization and visualization scheme to present extracted information components in a graphical form, facilitating to have a quick glance over the features and users' sentiments expressed over them without exploring the pile of review documents. The proposed scheme is capable to visualize mining results both from single as well as multiple review documents. It also provides a graphical interface for end users to explore and visualize statistical summary of feature-based opinions using bar and pie charts.