Name of the Candidate: **Tanvir Ahmad**
Name of the Supervisor: **Prof. Mohammad Najmud Doja**
Name of the Co-Supervisor: **Dr. Muhammad Abulaish**
Department: Dept. of Computer Engineering, F/O Engineering & Technology, J.M.I.
Topic: Frequent and Sequential Pattern Mining and their Applications

Abstract

In recent past, due to increasing popularity of the World Wide Web (WWW) and online social media, including online social networking sites, microblogging sites, online discussion forums, newsgroups, review sites, blogs, etc. there has been an unabated growth in user-generated contents causing the problem of *information overload*. Though the amount of useful information and knowledge contained in such data sources is very high, the research challenge lies in the fact that distillation of knowledge from such repository is very difficult, as most of them are either unstructured or semi-structured in nature. As a result, there is an increasing need of converting the information embedded within unstructured or semi-structured sources into a structured form, generally termed as *database curation*, without which the knowledge cannot be assimilated in a meaningful way which can be perceived by the users without exploring the pile of documents. Though, a number of techniques including document classification and clustering, information extraction, text summarization, etc. have been developed to analyze information contained in textual data, they are not sufficient to be applied on opinion data sources that have the intricacy of the embedded natural language in the documents. Though a good number of research efforts have been diverted towards analyzing opinion sources, including feature and opinion extraction, and sentiment analysis, to the best of our knowledge, no research effort has been made to identify sequential patterns among product features based on the associated opinions expressed by the users in such a way that the satisfaction/dissatisfaction of the user on a particular feature influence his/her satisfaction/dissatisfaction over other features.

In this thesis, we have proposed the design of a novel framework at the intersection of Information Retrieval and Extraction, Natural Language Processing, and Data Mining to mine frequent and sequential patterns from unstructured or semi-structured texts and use them for visualization and exploration of knowledge at different levels of granularity. The mined patterns are also used to design an imprecise query processing system to read users'

imprecise queries formulated through a guided manner and process them over textual data. The novelty of the work lies in the development of a **C**andidate **I**dentification and **F**requent **P**attern **G**eneration (CI-FPG) framework which exploits natural language processing and information extraction techniques to identify information components embedded within textual data and store them into a structural database which schema is derived in such a way that each constituent of the information components constitute an attribute. The CI-FPG framework is integrated with the FP-Growth algorithm, to mine frequent patterns from the generated structured database.

To analyze mined frequent patterns at different levels of granularity, a clustering method is proposed which customizes K-Means algorithm to classify features based on the opinions expressed over them. We have also proposed a ranking mechanism of features based on the sentiment scores of the associated opinions. To provide a comprehendible visualization of the mined patterns, we have proposed a feature-based cloud generation mechanism which helps users to get a visual depiction of the whole mining results. The proposed cloud generation method follows a two-layered star-shaped architecture, in which the product is placed at the central position, feasible features are placed at the first layer and connected to the central node, and finally the opinions are placed at second layer and connected to the respective features at first layer. We have also proposed a sequential pattern mining approach that can provide an added advantage to the users, especially the manufacturers, to know the sequence of features, with respect to underlying reviews, in which each feature has a ripple effect on other following features to influence users' sentiment about the product. Finally, we have proposed an application of mined frequent patterns to answer opinion-based imprecise queries formulated in a guided manner at different levels of granularity. The proposed query processing system uses well-known Bandler-Kohout Fuzzy Information Retrieval Model (BK-FIRM) which facilitates the processing of imprecise queries using fuzzy qualifiers and ranks the retrieved documents based on the degree of their relevance to the query. The efficacy of the proposed methods in this thesis is established through experimentation over datasets taken from various domains, including *digital camera* and *hotel*.