別紙4（課程博士）

Date of Submission:
平成　2016 年 09 月 23 日

Department 情報・知能工学専攻	Student ID Number 学籍番号	第 115309 号	Supervisors 指導教員	MASAKI AONO
Applicant's name 氏名	MD ZIA ULLAH			

Abstract
論文内容の要旨（博士）

Title of Thesis 博士学位論文名	Bipartite Graph Based Ranking Methods for Subtopic Mining and Genetic Disease Prediction (サブトピック・マイニングと遺伝病予測のための二部グラフに基づくランキング手法)

(Approx. 800 words)
(要旨　1,200 字程度)

With the vast amount of information available on the Internet in the forms of Web pages such as news articles, microblog posts, and shopping sites, a search engine has become an essential tool of our daily life to explore information on the Internet. When an information need comes up in mind, the user expresses it into a set of words (aka. a query) and issues the query to the search engine. Currently, given a query, a search engine responds a ranked list of documents to satisfy the information needs of the user. However, if the user's issued query conveys a variety of interpretations, the search result is far from the "what the user really wants to search." Therefore, we assume that "what the user really wants to search" is the user's "search intent."

According to user search behavior analysis, the search query is usually short, ambiguous, or may entail multiple search intents. Issuing the same query, users may have different information needs, which corresponds to diverse search intents. Traditional information retrieval models, including the boolean model and the vector space model, treat the issued query as a clear, well-defined representation, and completely neglect any sort of ambiguities. Ignoring the user's intents underlying a query, information retrieval models may result in documents, possibly containing too much relevant information on a particular aspect of a query. As these documents cover only a few intents or interpretations, the user may not be satisfied.

To satisfy the users' intents in their Web search, a practical approach is to diversify the documents for the given search query, that is to present a ranked list of documents by taking into account the coverage, popularity, and novelty of the search intents underlying a query. Therefore, exploring the possible search intents of the query is an essential need for the next-generation search engine.

Exploring the search intents underlying a query has gained much interest in recent years. Researchers have proposed several methods for mining subtopics as search intents by exploiting different resources, including the top retrieved documents, query logs, Wikipedia, anchor texts, and the query suggestions provided by the commercial search engines. Query suggestions hold some search intents, however, suggested queries are often noisy and possess a group of similar suggestions covering a single intent of the query. Moreover, the search query and the search intents (i.e. subtopics) are short in length. Thus, it is a challenging task to estimate the semantic and contextual similarity between a pair of short texts.

In this dissertation, we have developed a novel framework that explores the subtopics covering intents underlying a query, estimates subtopic importance, and diversify them by considering the relevance and novelty. To diversify the search results, we have devised a new way of ranking based on a new novelty estimation that faithfully represents the possible search intents of the query. For representing subtopic, we have proposed new semantic features based on a word-embedding model to capture the semantic matching of a query with a candidate subtopic. To rank a set of candidates, we have developed a bipartite graph-based ranking method of estimating the global importance of the candidate subtopic by aggregating the local importance of each feature.

Estimating the contextual similarity between a pair of short texts is a formidable task. Two short texts might not be lexically similar, however, semantically similar. Our observation is that if two short texts represent the similar meaning, even though they are not lexically similar, they may result in similar kinds of documents from a search engine. Mutual information between two probability distributions of words, extracted from the corresponding documents, may represent the contextual similarity between two short texts. Therefore, we have proposed a contextual similarity function for short texts through the probability distributions of terms in the top retrieved documents from a search engine.

We have experimented and evaluated the proposed methods, and compared with the earlier methods on benchmark data sets. We have conducted experiments on the intent mining test collections, including NTCIR-10 and NTCIR-12, and web corpus, including Clueweb09-Cat-B and Clueweb12-B13. Experimental results demonstrate the effectiveness of our proposed methods in comparison to the known earlier methods.

In the meantime, with a vast amount of medical knowledge available on the Internet, it is becoming increasingly vital to help doctors in clinical diagnostics by suggesting plausible diseases predicted with data and text mining technologies. In this dissertation, we have also proposed to rank genetic diseases for a set of clinical phenotypes. In this regard, we have associated a phenotype-gene bipartite graph (PGBG) with a gene-disease bipartite graph (GDBG) by producing a phenotype-disease bipartite graph (PDBG). To estimate the importance weight of an edge in PDBG, we have developed a Bidirectionally-induced Importance Weight (BIW) prediction method to PDBG by considering the content and link information from both sides of the bipartite graph. The experimental results exhibit that our proposed BIW method has outperformed the known previous methods in the disease retrieval system.