

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)

September, 2016

Doctor of Engineering

Md Zia Ullah

エムヂイ ジア ウンラ

Toyohashi University of Technology

Acknowledgements

Many organizations, including Japan's Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan Society for the Promotion of Science (JSPS), Japan Student Services Organization (JASSO), Amano Institute of Technology, and The Hori Sciences and Arts Foundation have funded the research presented in this thesis through scholarship or grant. I am grateful to them for their financial support.

Words will never be enough to express my gratitude to the best mentor in my life, my supervisor, Professor Masaki Aono for his spontaneous guidance, patience, and support throughout the years. Accepting me for a Master's program and then a Ph.D. program in Knowledge Data Engineering (KDE) laboratory, he guided my scientific and personal development for the last five years. His constant directions and guidelines were of inestimable value in polishing my research skills, and, consequently, overcoming all difficulties, which arose in these academic years. Moreover, I enjoyed his positive attitude to setup a nice working environment. Certifying him as the best academic scholar, I have ever experienced.

Expressing special gratitude to Dr. Md Hanif Seddiqui, Professor of the University of Chittagong, for his recommendation to my supervisor. I would like to appreciate the external reviewers of this thesis, Professor Shigeru Masuyama and Professor Kyoji Umemura, for their careful reading, invaluable comments, and suggestions. For sincere support, insightful comments, and suggestions during my presentations, I am indebted to Dr. Atsushi Tatsuma, Assistant Professor of KDE laboratory. Recalling all the teachers from the

elementary school to the current university, who polished my brain and imposed me, to become today's me. I would like to express my gratitude to all of my past and present lab mates in the KDE laboratory for their supportive mentality and continuous wishes. My heartiest gratitude goes to all the academic and administrative staff at Toyohashi University of Technology (TUT), who always supplies any information and reply my queries sincerely with a smiley expression.

Staying a long distance from home, these years could be more challenging without the support from my friends, the list is too long to mention here; my heartfelt affection goes to them. I also would like to express my thanks to all the people who I met during these years, and who made this long journey more enjoyable. Finally, I would like to express my respect, honor, and affection for my family members for their infinite love and constant encouragement to carry out my goals.

Abstract

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 "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.

Contents

Nomenclature	xii
1 Introduction	1
1.1 Background	1
1.2 Dissertation Focus	3
1.3 Our Contributions	4
1.4 Organization	5
2 Related Work	7
2.1 Search Intent Mining	7
2.2 Search Result Diversification	9
2.3 Bipartite Graph-based Ranking to Genetic Disease	10
2.4 Word Embedding	11
3 General Concepts and Terminology	12
3.1 Bipartite Graph	12
3.2 Word Embedding	13
3.3 Query Dependent Feature	16
3.3.1 TF-IDF	16
3.3.2 BM25	18
3.3.3 Divergence from randomness (DFR)	18
3.3.4 Language modeling with Dirichlet Smoothing	19
3.3.5 Language modeling with Jelinek-Mercer Smoothing	19
3.3.6 Term dependency Markov Random Field	20
3.3.7 Jensen-Shannon Divergence (<i>JSD</i>) based Similarity	20

3.3.8	Edit Distance based Similarity	21
3.3.9	Term Overlap	21
3.3.10	Term synonym overlap	21
3.3.11	Generalized Co-HITS	22
3.4	Query Independent Features	23
3.5	Ontology	24
4	Diversified Query Subtopic Mining	26
4.1	Introduction	26
4.2	Related Work	29
4.3	General Concepts and Terminologies	31
4.3.1	Search Intent	31
4.3.2	Subtopic	31
4.3.3	Subtopic Diversification	32
4.4	Diversified Subtopic Mining	32
4.4.1	Subtopic Candidate Generation	33
4.4.2	Subtopic Features Extraction	34
4.4.3	Subtopic Ranking	36
4.4.3.1	Supervised Feature Selection	37
4.4.3.2	Subtopic Relevance Estimation	38
4.4.3.3	Subtopic Diversification	41
4.5	Experiments and Discussion	43
4.5.1	Dataset	43
4.5.2	Evaluation Metrics	45
4.5.3	Experimental Settings	45
4.5.4	Important Features and Parameter Tuning	46
4.5.5	Experimental Results	49
4.5.6	Discussion	53
4.6	Summary	55

5	Bipartite Graph based Ranking of Genetic Disease	57
5.1	Introduction	58
5.2	Related Work	59
5.3	General Concepts and Terminologies	61
5.3.1	Phenotype	61
5.3.2	Genotype	62
5.3.3	Bipartite Graph (Bigraph)	62
5.3.4	Phenotype-Genotype Bipartite Graph	63
5.3.5	Gene-Disease Bipartite Graph	64
5.4	Methodology and Design	65
5.4.1	Data Acquisition	65
5.4.2	Proposed System Architecture	66
5.4.3	Exploring Causative Genes	66
5.4.4	Associating Bipartite Graphs	68
5.4.5	Estimating Candidate Weight	70
5.4.5.1	Bidirectionally-Induced Importance Weight (BIW) Method	71
5.4.5.2	TF-IDF Weight	73
5.4.5.3	BM25 Weight	73
5.4.5.4	Jensen-Shannon Divergence (JSD) Weight	75
5.4.5.5	Weighting Phenotype-Disease Bipartite Graph (PDBG)	75
5.4.6	Retrieving and Ranking the Diseases	75
5.5	Experiments and Evaluation	78
5.5.1	Query Set	78
5.5.2	Evaluation Methods	78
5.5.2.1	Kendall's Tau	79
5.5.2.2	Evaluation Setup	79
5.5.3	Comparison	80
5.6	Summary	85

6	Conclusions and Future Directions	86
6.1	Conclusions	86
6.2	Future Directions	87
6.2.1	Resource based subtopic mining	87
6.2.2	Aspect oriented subtopic ranking	88
6.2.3	Hierarchical subtopic mining	88
6.2.4	Genetic Disease Ranking	88
6.2.5	Matching Diseases and Phenotypes Ontologies	89
7	Related Publications	90
	References	110

List of Figures

3.1	A simple bipartite graph	13
3.2	An architecture of continuous bag-of-words (CBOW) model . .	14
4.1	Illustration of a query, the possible search intents, and the representative subtopics corresponding to each search intent.	32
4.2	This figure demonstrates the diversified subtopics covering search intent.	33
4.3	Diversified subtopic mining flow	34
4.4	Bipartite graph based representation of subtopics and features .	39
4.5	A topic "grilling" from INTENT-2 dataset, which is labelled by its intents with probabilities and a set of subtopics under each intent	44
4.6	An empirical analysis of the parameter λ_1 in equation. (4.6) for our proposed method bipartite graph based <i>BGR</i> ranking on INTENT-2 dataset. The X-axis indicates the parameter λ_1 and the Y-axis indicates the corresponding scores of I-rec, D-nDCG, and D#-nDCG at the cutoff rank 10.	47
4.7	An empirical study of the parameter λ_2 in equation. (4.7). The X-axis indicates the parameter λ_2 and the Y-axis indicates the corresponding scores of I-rec, D-nDCG, and D#-nDCG at the cutoff rank 10.	48
4.8	Sensitivity analysis of the diversification parameter γ . The X-axis indicates the different values of the parameter γ and the Y-axis indicates the corresponding scores of I-rec, D-nDCG, and D#-nDCG at the cutoff rank 10.	48

5.1	The phenotype-gene bipartite graph ($\mathcal{P}\mathcal{G}\mathcal{B}\mathcal{G}$) with unit weight of edge. The red-colored genes are denoted as disease-causative.	63
5.2	The gene-disease bipartite graph ($\mathcal{G}\mathcal{D}\mathcal{B}\mathcal{G}$) with unit weight of edge.	64
5.3	Proposed system architecture	67
5.4	A PPIN and an extended PPIN with some explored causative genes in green.	68
5.5	An extended gene-disease bipartite graph ($\mathcal{E}\mathcal{G}\mathcal{D}\mathcal{B}\mathcal{G}$) with unit weight of edge. The green are the newly explored candidate-causative genes.	69
5.6	A sample of phenotype-disease bipartite graph ($\mathcal{P}\mathcal{D}\mathcal{B}\mathcal{G}$). This figure depicts how the weight between the phenotype p_3 and the disease d_4 is estimated. The phenotype p_3 is connected to the diseases $d_1, d_3, d_4,$ and d_n with frequencies 7, 5, 6, and 6. The disease d_4 is also connected to the phenotypes $p_2, p_3,$ and p_4 with frequencies 7, 6, and 3. The weight of the edge (p_3, d_4) is approximated based on the importance of the phenotype p_3 and the disease d_4 in Equation (5.1).	72
5.7	Empirical study of $BM25$ parameters k_1 and b	74
5.8	An implementation of our disease retrieval system	77
5.9	The comparison of our proposed method BJW with the baseline methods including $BM25, TF-IDF,$ and JSD ; (a) $NDCG@20$ metric, (b) $MAP@20$ metric	81
5.10	The comparison of our proposed method BJW with $Phenomizer$; (a) $NDCG@10$ metric, (b) $MAP@10$ metric, (c) $NDCG@20$ metric, and (d) $MAP@20$ metric	82
5.11	The comparison of our proposed method BJW with link analysis-based algorithm $Co-HITS$; (a) $NDCG@22$ metric, (b) $MAP@22$ metric	83

Chapter 1

Introduction

1.1 Background

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 the information need into a set of words (a.k.a., a query) and issues the query to the search engine. Currently, given a query, a search engine responds a ranked list of documents to fulfill the information needs of the user. If the issued query is clear and meaningful, such as, "the value of PI," user may be satisfied with the search result. However, in most of the cases, user's issued query conveys a variety of interpretations, and the search result is far from "what the user really wants to search." Therefore, we assume that "what the user really wants to search" is the user's "search intent." For example, given a single word query "apple," it may refer to the company "Apple Inc" or "the Apple fruit." Assuming that the "Disney action movie" that we saw in the airplane was interesting. Later, we forgot the title of the movie and wanted to make sure in the search engine. Even if a search query is "Disney action movie," the search result is too large and may not reach to the "what is really looking." This is the scenario of the present search engine, which does not include the user intents and the various interpretations of the search query.

According to user search behavior analysis, the search query is usually vague, ambiguous, or may entail multiple intents (Song *et al.* (2009); Spärck-Jones *et al.* (2007)). Issuing the same query, users may have different information needs, which corresponds to diverse search intents (Ren *et al.* (2015)). With an ambiguous query such as "full house," users may seek different interpretations, including "full house youtube," "full house korean drama," and "full house movie." With a broad query such as "t-test," users may be interested in different subtopics¹, including "t-test example," "steps of t-test," and "t-test p value." However, it is not clear which subtopic of a broad query is actually desirable for a user (Wang *et al.* (2013a)). Some intents of a search query are constantly popular; however, some others intents are time-dependent. For example, the query "US Open" is more likely to be target the tennis open in September and the golf tournament in June (Nguyen & Kanhabua (2014)).

Traditional information retrieval models, such as 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 search query, information retrieval models may result in top ranked documents, possibly containing too much relevant information on a particular aspect² of a query. As these documents cover a few subtopics or interpretations, the user may not be satisfied. To fulfill the information needs of the user, an information retrieval model should result in a ranked list of documents that are not only relevant to the popular intents, but also covers the diverse intents of the search query.

From the above background, through exploring the possible intents of the search query, there is an urgent need for the next-generation search engine, that is to develop a mechanism that makes use of the diverse interpretations. For this reason, given the redundant information and the ambiguous word, the aim is to develop an algorithm that can optimally combine a variety of interpretations underlying a query.

¹Subtopic is the more specialization of the query covering an intent

²Intent and Aspect are interchangeably used

1.2 Dissertation Focus

Exploring the subtopics covering intents underlying a query has gained much interest in recent years (Liu *et al.* (2014); Sakai *et al.* (2013a)). Several methods were proposed for mining subtopics by exploiting different resources, including the top ranked documents, anchor text, query logs, Wikipedia disambiguation pages, Freebase (Bollacker *et al.* (2008)), and the query suggestion provided by the commercial search engines (Santos *et al.* (2010a); Wang *et al.* (2013a,c)). Query suggestions hold some intents (Hu *et al.* (2015); Santos *et al.* (2010a)), however, suggested queries are often noisy and possess a group of similar suggestions covering a single intent of the search query. Moreover, the search query and the candidate subtopic are short in length. Therefore, it is a challenging task to estimate the similarity between a pair of short texts.

To combine the multiple intents of the search query in document retrieval, a sensible approach is to diversify the documents initially retrieved for the original search query. Diversification refers to the ranking of documents by taking into account the coverage, popularity, and novelty of the search intents underlying a query. However, diversification is formulated as an optimization problem that aims at optimizing an objective function with regard to the relevance and diversity (Agrawal *et al.* (2009); Carterette (2011); Santos *et al.* (2010a)). Moreover, it is an instance of the maximum coverage problem, a classical NP-hard problem in computational complexity theory.

In this dissertation, we have developed a novel framework that explores the subtopics covering intents underlying a search query, estimates subtopic relevance, and diversify them considering the relevance and redundancy. At the same time, to diversify the search results, we have proposed a new way of ranking based on a new novelty estimation that exploits the search intents of the query.

We have experimented and evaluated all the components of the proposed method, and compared the performance of our proposed method with the state-of-the-art intent mining methods on benchmark datasets. We have conducted experiments on the NTCIR intent mining and TREC web track diversity test

collections. We have utilized the document corpus, including *Clueweb09 Cat-B* and *Clueweb12 B13* for document retrieval. Moreover, a popular academic search engine, Indri (Strohman *et al.* (2005)) has been leveraged to index the corpus for baseline retrieval. Experimental results on the benchmark datasets demonstrate the effectiveness of our proposed methods in comparison to the known related works.

1.3 Our Contributions

The key contributions of this dissertation are summarized as follows:

1. **Word Embedding-based Features:** In order to capture the importance of the semantic matching of a query with a document, we propose three new semantic features based on the locally-trained word embedding model, including maximum word similarity (*MWS*), mean vector similarity (*MVS*), and uncommon word similarity (*UWS*) (ULLAH & AONO (2016); Ullah *et al.* (2016b)).
2. **Short Text Similarity:** 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 propose to estimate the contextual similarity for short texts, which is used to estimate the novelty for result diversification (ULLAH & AONO (2016); Ullah *et al.* (2016b)).
3. **Bipartite Graph-based Ranking:** We hypothesize that a relevant document should be ranked at the higher position by multiple effective features, and intuitively, an effective feature should be weighted higher by multiple relevant documents. Large weight might be given to a document that

tends to be ranked highly by a group of effective features, and vice versa. Therefore, there might be a weight propagations from features to documents and from documents to features. On these intuitions, we represent a set of features and a set of candidate documents as a bipartite graph, and introduce weight propagation from both sides of the bipartite graph. Given a set of features and a set of candidate documents, we propose a bipartite graph-based ranking (BGR) method to estimate the global importance of candidate documents by aggregating the local importance of the individual feature (ULLAH & AONO (2016); Ullah *et al.* (2016b)).

4. Associating Bipartite Graphs: Given two sets of bipartite graphs, we propose to associate one bipartite graph with another bipartite graph based on the transitive property among the nodes of bipartite graphs. By associating two bipartite graphs, all the information are embedded in a new bipartite graph (Ullah *et al.* (2013a, 2015)).
5. Estimating Weights of Edges in Bipartite Graph: In order to estimate the importance weight of an edge in a bipartite graph, we propose a Bidirectionally-induced Importance Weight (BIW) prediction method by considering content and link information from both sides of the bipartite graph (Ullah *et al.* (2015)).

1.4 Organization

The rest of the dissertation is organized as follows.

Chapter 2 discusses the related work.

Chapter 3 includes the general concepts and terminologies used throughout the dissertation to comprehend the readers about the contents of this dissertation.

Chapter 4 describes our proposed diversified subtopics mining method.

Chapter 5 describes our proposed bipartite graph based ranking of genetic disease.

Chapter 6 includes the conclusion and future directions of this dissertation.

1.4 Organization

Moreover, after each of the references in the reference section, we show the page number (in blue color) of this dissertation, where the reference was cited.

Chapter 2

Related Work

In this chapter, we conduct a literature review of previous work on search intent mining, search diversification, bipartite graph-based ranking, and word embedding.

2.1 Search Intent Mining

Web queries are usually short, ambiguous, and/or underspecified (Clarke *et al.* (2009); Song *et al.* (2009); Spärck-Jones *et al.* (2007)). To understand the meaning of queries, researchers define taxonomies and classify queries into predefined categories. Song *et al.* (2009) classified queries into three categories: ambiguous query, which has more than one meaning; board query, which covers a variety of subtopics; and clear query, which has a specific meaning or narrow topics.

At the query level, Broder (2002) divided query intent into navigational, informational, and transactional types. Nguyen & Kan (2007) classified queries into four general facets, including ambiguity, authority, temporal sensitivity, and spatial sensitivity. Boldi *et al.* (2008) created a query-flow graph with query phrase nodes and used them for query suggestion. Query suggestion is a key technique for generating alternative queries to help users drill down to a subtopic of the original query (Mei *et al.* (2008); Zhang & Nasraoui (2006)). In contrast to the query suggestion, subtopic mining focuses more on the diversity of possible subtopics of the original query rather than inferring relevant queries.

Hu *et al.* (2009) leveraged the knowledge contained in Wikipedia to predict the possible subtopics for a given query. Radlinski *et al.* (2010) proposed a method for inferring query intents from query reformulations and user click-through data. Santos *et al.* (2010a) exploited the query completions of a search engine to mine sub-queries (i.e. subtopics) for diversifying Web search results.

Wang *et al.* (2013a) proposed a method to mine subtopics of a query either directly from the query itself or indirectly from the top retrieved documents. In the direct approach, several external resources, such as Wikipedia, open directory project (*ODP*), query logs, and the related search services are investigated to mine subtopics. In indirect approach, subtopics are extracted by clustering, topic modeling, and concept-tagging of the top retrieved documents. The surrounding text of query terms in the top retrieved documents was also utilized to mine and rank subtopics (Wang *et al.* (2013c)). Recently, two-level hierarchical intents based search diversification methods were also proposed (Hu *et al.* (2015)).

Moreno *et al.* (2014) proposed an algorithm called Dual C-Means to cluster search results in dual-representation spaces with query logs and represented the cluster label as a subtopic. Damien *et al.* (2013) proposed a method for subtopic mining and ranking by fusing multiple resources. Kim & Lee (2015) proposed a frequent pattern-based method to mine candidate subtopics from a set of implicitly relevant documents.

Despite the fact that previous methods leveraged many resources to mine candidate subtopics, however, their ranking methods caused some noisy and redundant subtopics in the top rank. In contrast to previous methods, we introduce the locally-trained word embedding to extract semantic features and effectively diversify the candidate subtopics covering possible intents of the query by balancing the relevance and novelty.

NTCIR¹ has been organizing a research competition on *query subtopic mining* in Chinese, English and Japanese languages for the last couple of years, including NTCIR-10 INTENT-2¹, NTCIR-11 IMINE-1², and NTCIR-12 IMINE-2³.

¹<http://research.nii.ac.jp/ntcir/index-en.html>

¹<http://research.microsoft.com/en-us/projects/intent/>

²<http://www.thuir.org/IMine/>

³<http://www.dl.kuis.kyoto-u.ac.jp/imine2/>

Some approaches have been proposed by the participants exploiting multiple resources are discussed in (Kim & Lee (2013); Moreno & Dias (2013, 2016); Ullah *et al.* (2013b, 2016a); Wang *et al.* (2013b); Xue *et al.* (2013); Yue *et al.* (2016)).

2.2 Search Result Diversification

To satisfy general users in Web search, an information retrieval model should select a list of documents that are not only relevant to the popular intents, but also covers different intents of the search query. To do that, a sensible approach is to diversify the documents initially retrieved for the query (Clarke *et al.* (2008)) based on the search intents of the query. Search diversification is the process of ranking the documents initially retrieved for the query, taking into account the coverage, importance, and novelty of the documents with respect to search intents. However, search diversification is formulated as an optimization problem that aims at optimizing an objective function with regard to the relevance and diversity (Agrawal *et al.* (2009); Carterette (2011); Santos *et al.* (2010a)). It is an instance of the maximum coverage problem, a classical NP-hard problem in computational complexity theory.

Most diversification methods in the literature differ by how they implement the objective function either as an implicit or explicit approach. Maximal Marginal Relevance (MMR) (Carbonell & Goldstein (1998)) iteratively selects the best document by balancing the relevance with the query and novelty with other already selected documents in terms of cosine similarities in vocabulary. Many researchers proposed heuristic-based diversification methods by exploiting explicitly mined subtopics for the query. IA-Select proposed by Agrawal *et al.* (2009), an intent-aware diversifying method with topical categories of queries and documents based on ODP taxonomy. xQuAD proposed by Santos *et al.* (2010a), a greedy diversification algorithm to maximize the coverage of explicit mined query subtopics. PM2 proposed by Dang & Croft (2012), a framework for optimizing proportionality for result diversification, which is motivated by the problem of assigning seats to members of competing political parties. Yu & Ren (2014) formulated diversification as a 0-1 multiple subtopic knapsack problem. Fusion diversification proposed by Liang *et al.* (2014), inferred latent subtopics

2.3 Bipartite Graph-based Ranking to Genetic Disease

based on topic modeling. Some researchers use machine learning techniques to diversify search results, including Structural SVMs (Yue & Joachims (2008)) and R-LTR (Zhu *et al.* (2014)). Recently, Xia *et al.* (2016) proposed to model the novelty of document based on the neural tensor network for search research diversification.

NTCIR has been organizing a research competition on *search result diversification* in Chinese, English, and Japanese languages for the last couple of years, including NTCIR-11 IMINE-1¹, and NTCIR-12 IMINE-2². TREC³ has organized web track for diversified retrieval evaluation, including Web track 2009, 2010, 2011, and 2012.

2.3 Bipartite Graph-based Ranking to Genetic Disease

Many real applications can be modeled as bipartite graph, such as Viewers and Movies in a movie recommendation system (Bogers (2010)), Video shots and Tags (YANAI *et al.* (2015)) in video tagging system, Traders and Stocks (Sun *et al.* (2005)) in a financial trading system, Authors and Papers in a scientific literature (Barabási *et al.* (2002)), Conferences and Authors in a scientific publications network (Wang *et al.* (2013d)), Queries and URLs in query logs (Deng *et al.* (2009a)), Entities and Co-List (Cao *et al.* (2011)) in a Web page for entity ranking, Phenotypes and Diseases (Ullah *et al.* (2015)) in a disease prediction system, etc. Similarly, we represent a set of features and a set of candidate documents as a bipartite graph, and propose a bipartite graph-based ranking (*BGR*) method to estimate the global importance of the candidate documents by aggregating the local importance of each feature.

In the postgenomic era, it is widely established in Bioinformatics and molecular biology to represent the associations between biomedical entities as networks, and to analyze their topology to obtain a global understanding of underlying relationships (Barabási *et al.* (2011); Butts (2009); Yıldırım *et al.* (2007)). In

¹<http://www.thuir.org/IMine/>

²<http://www.dl.kuis.kyoto-u.ac.jp/imine2/>

³<http://trec.nist.gov/data/webmain.html>

this regard, *DisGeNET* is a coherent tool that analyzes and interprets the human gene network to disease network (Bauer-Mehren *et al.* (2010)). It visualizes the gene-disease association network as a bipartite graph and provides gene-centric and disease-centric views of the data. Another system which infers the genotype-phenotype bipartite relationship using the random walk with restart algorithm to the heterogeneous network (*RWRH*) (Li & Patra (2010)), where a heterogeneous network is constructed by connecting the gene network and phenotype network using the phenotype-gene bipartite relationship from the OMIM database.

2.4 Word Embedding

Word embedding is a real-valued vector representation of words. The idea behind the word embedding is to map the whole vocabulary of a language into a vector space in such a way that, words that are semantically and syntactically similar tend to be close in this embedding space. The similarity of two words can be quantified by the similarity of their embedding vectors. Neural network-based method, *word2vec* was proposed by Mikolov *et al.* (2013a,b) have been shown to be surprisingly effective at capturing the semantics, which is useful for various Natural Language Processing (NLP) and reasoning tasks, including word analogies. Inspired by the effectiveness of word embedding in NLP, recently, word embedding has also been applied in information retrieval contexts, including term re-weighting (Zheng & Callan (2015)), short-text similarity (Kenter & de Rijke (2015)), query expansion (Zamani & Croft (2016)), and estimating generalized language model (Ganguly *et al.* (2015)).

Chapter 3

General Concepts and Terminology

This chapter introduces some of the basic definitions to familiarize the reader with the notions and terminologies used throughout the dissertation. It describes about the bipartite graph, word embedding, query-dependent features, query-independent features, and ontology.

3.1 Bipartite Graph

A bipartite graph, also called a bigraph, is a set of graph vertices decomposed into two disjoint sets such that no two graph vertices within the same set are adjacent. Bipartite graphs are equivalent to two-colorable graphs, and a graph is bipartite if and only if all its cycles are of even length.

Consider a bipartite graph $\mathcal{G} = (\mathcal{U} \cup \mathcal{V}, \mathcal{E})$, its vertices can be divided into two disjoint sets \mathcal{U} and \mathcal{V} such that each edge in \mathcal{E} connects a vertex in \mathcal{U} and one in \mathcal{V} ; that is, there is no edge between two vertices in the same set. A bipartite graph is depicted in Figure 3.1.

Definition 1 *A bipartite graph is a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ whose vertex set \mathcal{V} can be partitioned into two non empty sets \mathcal{V}_1 and \mathcal{V}_2 in such a way that every edge in \mathcal{E} of \mathcal{G} joins a vertex in \mathcal{V}_1 to a vertex in \mathcal{V}_2 .*

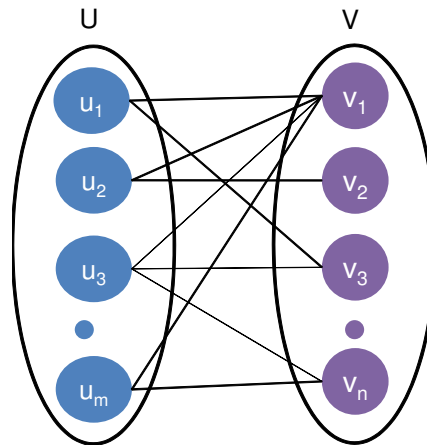


Figure 3.1: A simple bipartite graph

3.2 Word Embedding

Word embedding is defined as the real-valued vector representation of words. The idea behind the word embedding is to map the whole vocabulary of a language into a vector space such that words that are semantically and syntactically similar, tend to be close in this embedding space. Neural network-based method, *word2vec* which was proposed by Mikolov *et al.* (2013a,b) introduced two architectures for word embedding, including Continuous Bag-of-Words (CBOW) and Skip-Gram. We will describe the CBOW model, although our proposed word embedding based features, described in Chapter 4, are also applicable to vectors produced by Skip-Gram model.

A word embedding $F : words \rightarrow \mathbb{R}^p$ is a parameterized function, which maps words to high-dimensional vectors (i.e. p is 50 to 1000). For example, the word vectors corresponding to “iPhone” and “Apple” are shown as follows:

$$W(\text{“iphone”}) = (\dots, 0.2, \dots, -0.4, \dots, 0.7, \dots)$$

$$W(\text{“apple”}) = (\dots, 0.0, \dots, 0.6, \dots, -0.11, \dots)$$

Let us consider a context, $(x_{t-n}, \dots, x_{t-1}, x_{t+1}, \dots, x_{t+n})$, consisting of multiple words within a fixed-sized window around a target word x_t . We illustrate a schematic diagram of the CBOW model in Figure 3.2, where V denotes the number of words in the vocabulary, N denotes the number of units in the

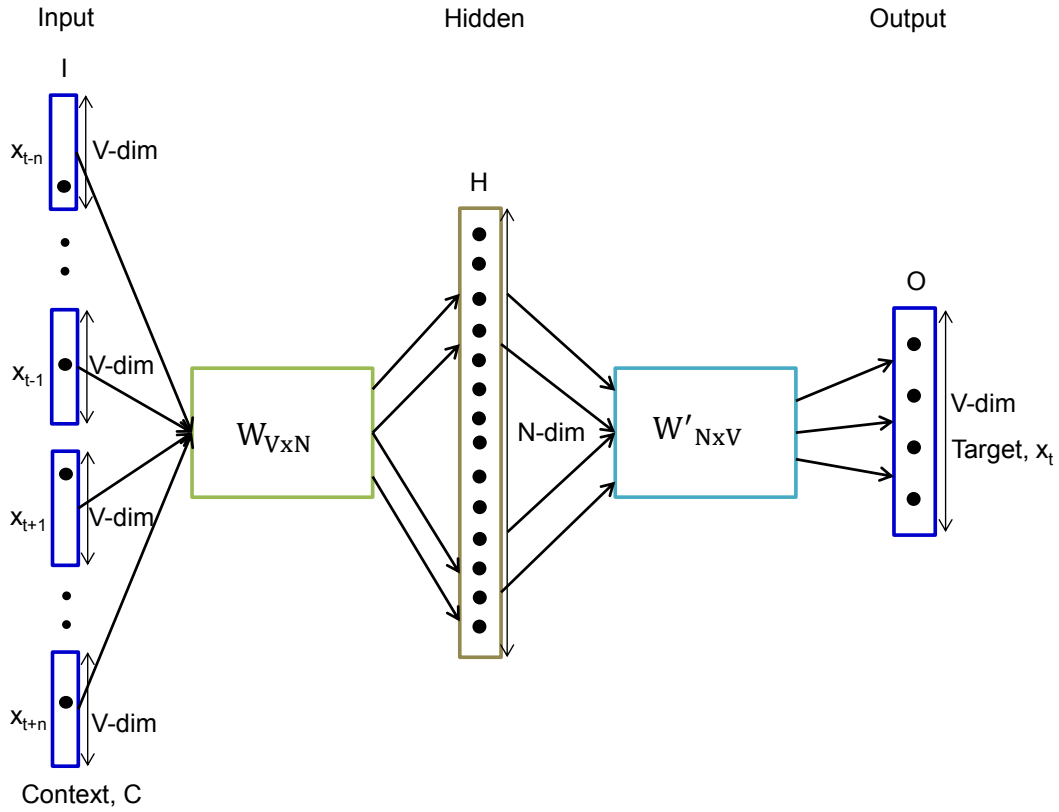


Figure 3.2: An architecture of continuous bag-of-words (CBOW) model

hidden layer (H), $W_{V \times N}$ denotes the input matrix from the input (I) to hidden layer (H), $W'_{N \times V}$ denotes the output matrix from the hidden (H) to output layer (O), and C denotes the number of words in the context. Both the input matrix W and the output matrix W' are initialized randomly.

Each word in the context is represented with V -dimension as a one-hot encoded vector, that means, only one out of V units, $\{x_1, \dots, x_V\}$, will be 1, and all other units are 0 (Rong (2014))). All one-hot encoded vectors of the context words are given to CBOW model as input and the input matrix W is shared by all context words.

Given the vectors of the context words as input, the vector of hidden layer h is linearly computed by averaging vectors of the product of input matrix W

and the vectors of context words as follows:

$$\begin{aligned} \mathbf{h} &= \frac{1}{C} W^T \cdot (\mathbf{x}_{t-n} + \cdots + \mathbf{x}_{t-1} + \mathbf{x}_{t+1} + \cdots + \mathbf{x}_{t+n}) \\ &= \frac{1}{C} (\mathbf{v}_{w_{t-n}}^T + \cdots + \mathbf{v}_{w_{t-1}}^T + \mathbf{v}_{w_{t+1}}^T + \cdots + \mathbf{v}_{w_{t+n}}^T) \end{aligned} \quad (3.1)$$

where $(\mathbf{x}_{t-n}, \cdots, \mathbf{x}_{t-1}, \mathbf{x}_{t+1}, \cdots, \mathbf{x}_{t+n})$ are the one hot-encoded vectors of the context words and \mathbf{v}_{w_i} is the i -th row vector of input matrix W .

Given the output matrix W' and the vector of hidden layer \mathbf{h} , we can compute the input to each unit in the output layer (O) as follows:

$$u_j = \mathbf{v}'_{w_j}{}^T \cdot \mathbf{h} \quad (3.2)$$

where u_j is the input to the j -th unit in the output layer (O). $\mathbf{v}'_{w_j}{}^T$ denotes the j -th column in the output matrix W' .

We can estimate the output score of each unit in the output layer (O) by passing its input score from Equation 3.2 through a soft-max function. The output of the j -th unit in the output layer (O), y_j is computed as follows:

$$y_j = p(x_{y_j} | x_{t-n}, \cdots, x_{t-1}, x_{t+1}, \cdots, x_{t+n}) = \frac{e^{u_j}}{\sum_{j'=1}^V e^{u_{j'}}} \quad (3.3)$$

By substituting the Equation 3.2 in the Equation 3.3, we may have the following equation:

$$y_j = p(x_{y_j} | x_{t-n}, \cdots, x_{t-1}, x_{t+1}, \cdots, x_{t+n}) = \frac{e^{\mathbf{v}'_{w_j}{}^T \cdot \mathbf{h}}}{\sum_{j'=1}^V e^{\mathbf{v}'_{w_{j'}}{}^T \cdot \mathbf{h}}} \quad (3.4)$$

The objective function of CBOW is to maximize the log conditional probability of the target word x_t given the context words. Therefore, the loss function is

defined as follows:

$$\begin{aligned}
 E &= -\log p(x_t | x_{t-n}, \dots, x_{t-1}, x_{t+1}, \dots, x_{t+n}) \\
 &= -u_{j^*} + \log \sum_{j'=1}^V e^{u_{j'}} \\
 &= -\mathbf{v}'_{w_t} \cdot \mathbf{h} + \log \sum_{j'=1}^V e^{\mathbf{v}'_{w_{j'}} \cdot \mathbf{h}}
 \end{aligned} \tag{3.5}$$

where j^* is the index of the actual target word x_t .

The weights of input matrix W and output matrix W' are learned using back propagation. To learn weight matrices W and W' , the training examples are fed into the *CBOW* model. The prediction error is observed through the difference between the predicted output and the actual output. Then, with respect to the elements of both output matrix W' and input matrix W , the gradients of this prediction error are estimated. Both of the matrices are corrected in the direction of these gradients. The row vectors of the input matrix W or the column vectors of the output matrix W' are the corresponding vectors of words in the vocabulary V .

3.3 Query Dependent Feature

Let Q be a query and $\mathcal{D} = \{D_1, D_2, D_3, \dots, D_N\}$ be a set of documents. Query-dependent features are directly computed by scoring the occurrences of the terms of the query Q in each document D of \mathcal{D} . Among the query-dependent features, there are some term frequency, language modeling, term dependency, lexical, and mutual information based features.

3.3.1 TF-IDF

The *TF-IDF* is a basic technique often used in information retrieval and text mining (Salton *et al.* (1983)). This is a statistical measure used to evaluate how important a term is to a document in a collection or corpus.

3.3 Query Dependent Feature

Term Frequency (*TF*) is defined as the count of the term in a document divided by the total number of terms in it. It is usually normalized to prevent a bias toward longer documents, which may have a higher term frequency regardless of the actual importance of that term in the document, to give a measure of the importance of the term t_i within the particular document d_j , where i and j are the indices of term and document, respectively. Therefore, the term frequency is defined as follows:

$$TF_{t_i, d_j} = \frac{|t_i \in d_j|}{\sum_k |t_k \in d_j|}$$

where $|t_i \in d_j|$ is the number of occurrences of the i^{th} term, t_i in the j^{th} document, d_j , and the denominator $\sum_k |t_k \in d_j|$ is the sum of the number of occurrences of all terms t_k in the j^{th} document d_j , where k varies from 1 to the number of distinct terms in document d_j , that is, the size of the document $|d_j|$.

The Inverse Document Frequency (*IDF*) is a measure of the general importance of the term of the ratio of the total number of documents to the number of documents containing the term, and then taking the logarithm of that ratio.

$$IDF_{t_i, D} = \log \frac{|D|}{|\{d \in D | t_i \in d\}|}$$

where $|D|$ is the total number of documents in the corpus, $|\{d \in D | t_i \in d\}|$, is the number of documents where the i^{th} term t_i appears (that is $N_{t_i, d_j} \neq 0$). If the term is not in the corpus, this will lead to a division-by-zero. It is therefore common to use $1 + |\{d \in D | t_i \in d\}|$.

The overall relevancy of a document with respect to a term can be computed using both the TF and IDF. Therefore, the weight of a term t_i in a document d_j is defined as follows:

$$weight(t_i, d_j) = TF \cdot IDF_{t_i, d_j} = TF_{t_i, d_j} \times IDF_{t_i, D}$$

This measure is called *TF·IDF* weight. A document can be considered as a multi-dimensional vector, where each dimension represents a term with the *TF·IDF* as its weight.

3.3.2 BM25

The Okapi best matching 25 (*BM25*) (Sparck Jones *et al.* (2000)) approach was based on the probabilistic retrieval framework developed in the 1970s and 1980s by (Robertson & Zaragoza (2009)) (1981). The *BM25* formula is used for measuring the similarity between a user query q and a document d . It is used to rank a set of documents based on the query terms appearing in each document, regardless of the inter-relationship between the query terms within a document (e.g., their relative proximity). It is not a single function, but actually a whole family of scoring functions, with slightly different components and parameters. One of the most prominent instantiations of the function is as follows. Given a query Q , containing a set of keywords $\{q_1, q_2, \dots, q_n\}$, the *BM25* score of a document D for the query Q is defined as follows:

$$weight(Q, D) = \sum_{i=1}^n \frac{TF_{q_i, D} \cdot (k_1 + 1)}{k_1 \cdot ((1 - b) + (b \cdot \frac{|D|}{avg_l})) + TF_{q_i, D}} \times \log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5}$$

where $TF_{q_i, D}$ is the q_i 's term frequency in the document D , N is the total number of documents in the collection, $\frac{|D|}{avg_l}$ is the ratio of the length of document d to the average document length, and $n(q_i)$ is the number of documents, where the term q_i appears. k_1 and b are free parameters, usually chosen, in the absence of an advanced optimization, as $k_1 \in [1.2, 2.0]$ and $b = 0.75[1]$.

3.3.3 Divergence from randomness (DFR)

A non-parametric divergence from randomness (DFR) based models, DFH (Amati *et al.* (2008)) has been shown to perform effectively across a variety of Web search tasks (Santos *et al.* (2010b)). Term frequency based DFH feature, $f_{DFH}(Q, D)$ is computed as:

$$f_{DFH}(Q, D) = \sum_{w \in Q} \frac{tf_{w, D} (1 - \frac{tf_{w, D}}{l_D})^2}{tf_{w, D} + 1} \log_2 (tf_{w, D} \frac{avg_l_D}{l_D tf_{w, C}}) + 0.5 \log_2 (2\pi tf_{w, D} (1 - \frac{tf_{w, D}}{l_D})) \quad (3.6)$$

where $tf_{w,D}$ is the frequency of the term w in the document D and $tf_{w,C}$ is the frequency of the term in the collection C .

3.3.4 Language modeling with Dirichlet Smoothing

Language modeling with Dirichlet smoothing feature, $f_{LMDS}(Q, D)$ is a language modeling approach to information retrieval (Song & Croft (1999)) and smoothed using Dirichlet smoothing (Zhai & Lafferty (2001)). It is computed as the log likelihood of the query being generated from the document. The feature is computed as:

$$f_{LMDS}(Q, D) = \sum_{w \in Q} tf_{w,Q} \log \frac{tf_{w,D} + \mu P(w|C)}{|D| + \mu} \quad (3.7)$$

where $tf_{w,Q}$ is the number of times that w occurs in the query, $tf_{w,D}$ is the number of times w occurs in the document D , $|D|$ is the number of terms in the document, $P(w|C)$ is the background language model, and μ is a tunable smoothing parameter.

3.3.5 Language modeling with Jelinek-Mercer Smoothing

Language modeling with Jelinek-Mercer smoothing (Zhai & Lafferty (2001)) feature, $f_{LMJM}(Q, D)$ is defined as the linear combination of the probability of the query term given the document and the probability of the query term in background language model, and computed as:

$$f_{LMJM}(Q, D) = \sum_{w \in Q} \lambda P(w|D) + (1 - \lambda) P(w|C) \quad (3.8)$$

where $P(w|D)$ is the probability of the query term w given the document D , $P(w|C)$ is the background language model, and λ is the controlling parameter. The parameter λ is set to $\frac{\mu}{|S| + \mu}$, where the parameter μ is set to 2,500.

3.3.6 Term dependency Markov Random Field

A particularly effective approach to exploit term dependency was proposed by Metzler & Croft (2005). In this model, unigram, bigram sequential dependency, and bigram full dependency is linearly interpolated. The term dependency with Markov random field based feature, $f_{MRF}(Q, D)$ is computed as:

$$\begin{aligned}
 f_{MRF}(Q, D) = & \alpha_u \sum_{w_i \in Q} \log P(w_i | \theta_D) + \\
 & \alpha_s \sum_{w_i \in Q} \sum_{\substack{w_j \in Q \\ j=i+1}} \log P(\langle w_i, w_j \rangle | \theta_D) + \\
 & \alpha_f \sum_{w_i \in Q} \sum_{\substack{w_j \in Q \\ j \neq i}} \log P(\langle w_i, w_j \rangle | \theta_D)
 \end{aligned} \tag{3.9}$$

where the parameters α_u , α_s , and α_f control the weights of the unigram, bigram sequential, and bigram full models in the linear combination, respectively, and θ_D is the language model of the document D . Here, α_u , α_s , and α_f are the free parameters, which reflect the importance of each component.

3.3.7 Jensen-Shannon Divergence (JSD) based Similarity

Consider the set $M_+^1(A)$ of probability distributions, where A is a set provided with some σ -algebra of measurable subsets. In particular, we can take A to be a finite or countable set with all subsets being measurable. The Jensen-Shannon divergence (JSD) (Lin (1991)): $M_+^1(A) \times M_+^1(A) \rightarrow [0, \infty)$ is a symmetrized and smoothed version of the Kullback-Leibler divergence $D(P \parallel Q)$. It is defined by

$$JSD(P \parallel Q) = \frac{1}{2}D(P \parallel M) + \frac{1}{2}D(Q \parallel M)$$

where $M = \frac{1}{2}(P + Q)$

The distance between two probability distributions P and Q can be calculated using the above formula, $JSD(P \parallel Q)$. To compute the similarity, the weight is defined as follows:

$$weight(P|Q) = 1.0 - JSD(P|Q)$$

3.3.8 Edit Distance based Similarity

To measure the lexical similarity between the query and the document, edit distance (Shi & Yang (2006)) based feature, $f_{EDS}(Q, D)$ is computed as:

$$f_{EDS}(Q, D) = 1 - \frac{Edit_distance(Q, D)}{Max(length(Q), length(D))} \quad (3.10)$$

where *Edit_distance* is measured based on the number of edit operations (insertion, deletion, or substitution of a word) necessary to unify two strings, and $length(Q)$ is the number of terms in the query Q .

3.3.9 Term Overlap

Term overlap feature, $f_{Overlap}(Q, D)$ is simply defined as the fraction of query terms that occur, after stemming and stopping, in the candidate document (Metzler & Kanungo (2008)). It is computed as:

$$f_{Overlap}(Q, D) = \frac{\sum_{w \in Q} I(w \in D)}{|Q|} \quad (3.11)$$

3.3.10 Term synonym overlap

Term synonym overlap feature, $f_{Overlap-syn}(Q, D)$ is the generalization of the overlap feature by considering synonyms of query terms. It is defined as the fraction of query terms that either match with document term or have a synonym that matches with document term. It is computed as:

$$f_{Overlap-syn}(Q, D) = \frac{\sum_{w \in Q} I(Syn(w) \in D)}{|Q|} \quad (3.12)$$

where the function, $Syn(w)$ returns the synonyms of the term w . WordNet 3.0 (Fellbaum (1998)) is to determine the synonyms of noun, adjective, verb, and adverb POS, followed by canonicalization with Krovetz stemmer (Krovetz (1993)) and POS tagging with Stanford NLP parser (Socher *et al.* (2013)).

3.3.11 Generalized Co-HITS

Co-HITS is a generalized link analysis algorithm for bipartite graph (Deng *et al.* (2009a)). It incorporates the bipartite graph with the content information from both sides as well as the constraints of relevance, and investigated based on iterative and regularization framework. The basic idea of the iterative framework is to propagate the scores on the bipartite graph via an iterative process with the constraints from both sides, where it contains *HITS*, personalized page rank, and one step propagation as a special case.

Given a query q , content information, and the matrix representation of the bipartite graph, the ultimate goal of this algorithm is to find a set of entities which are most relevant to the query q . Consider a bipartite graph $G = (U \cup V, E)$, where U and V are two disjoint sets such that every *edge* in E connects a vertex in U to one in V . Let W^{UV} denote the transition matrix from U to V and W^{VU} denote the transition matrix from V to U , where each entry contains a weight $w_{u_i v_j}$ from vertex u_i to v_j . To consider the vertices in one side, then the hidden transitional matrix in set U can be introduced as W^{UU} , where each entry $w_{u_i u_j} = \sum_{k \in V} w_{u_i v_k} w_{v_k u_j}$. Therefore, we can easily obtain the transition matrices W^{UV} , W^{VU} , W^{UU} , and W^{VV} .

The basic idea of the *Co-HITS* algorithm is to propagate scores on the bipartite graph via an iterative process. Let $u_i \in U$ and $v_k \in V$ be two vertices in the bipartite graph, and x_i is the score of u_i and y_k is the score of v_k at some state. The score y_k of vertex $v_k \in V$ is propagated to the vertex $u_i \in V$ according to the transition probability. Similarly, additional scores are propagated from other vertices of V to u_i , then the score x_i of u_i is updated to get a new value x_i . Similarly, the new value x_i is propagated to v_k . The intuition behind the score propagation is the mutual reinforcement to boost co-linked entities in the bipartite graph. In addition, the initial relevance scores based on the content information provide invaluable information.

In order to incorporate the bipartite graph with content information, the generalized *Co-HITS* equations can be written as follows:

$$x_i = (1 - \lambda_u)x_i^0 + \lambda_u(1 - \lambda_v) \sum_{k \in V} w_{ki}^{vu} y_k^0 + \lambda_u \lambda_v \sum_{j \in U} w_{ji}^{uu} x_j \quad (3.13)$$

$$y_j = (1 - \lambda_v)y_j^0 + \lambda_v(1 - \lambda_U) \sum_{k \in U} w_{kj}^{uv} x_k^0 + \lambda_u \lambda_v \sum_{k \in V} w_{ki}^{vv} y_k \quad (3.14)$$

where $\lambda_u \in [0,1]$ and $\lambda_v \in [0,1]$ are the personalized parameters, x_i^0 and y_k^0 are the initial scores for u_i and v_k respectively.

The final score x_i of vertex u_i can be obtained through an iteratively updating process by using Equation 3.13, and the final score y_j of vertex v_j can be obtained through an iteratively updating process by using Equation 3.14. The first term of the Equation 3.13 or 3.14 is the initial score, the second term is the one-step propagation and the third term is the personalized PageRank.

3.4 Query Independent Features

The goal of query independent features is to encode a prior knowledge that we may have about individual document. We describe some simple query independent features as follows.

Given a query Q , Voting feature, $f_{Voting}(D)$ is defined as the number of search engines, which returns a document D .

$$f_{Voting}(D) = \sum_{i=1}^N I(D \in R_i) \quad (3.15)$$

where I is an indicator function, which returns 1 if its argument is true, N is the number of Search engines, R_i is the ranked list of documents from the Search engine i .

Reciprocal rank feature, $f_{RR}(D)$ is defined as the reciprocal of the rank of the document in the search results for the query Q in across search engines. It is computed as:

$$f_{RR}(D) = \sum_{i=1}^N \frac{1}{\sqrt{rank_i(D) + 1}} \quad (3.16)$$

where $rank_i(D)$ is the rank of the document in the i^{th} search engine and N is the number of Search engines.

Longer terms in the document would reflect a more thoughtful and readable style. To focus on the readability of the document, average term length (ATL) in a document is defined as:

$$f_{ATL}(D) = \frac{1}{l_D} \sum_{w \in D} t f_{w,D} l_w \quad (3.17)$$

where l_w denotes the length of the term w in character.

Additional readability features have been recently proposed is topic cohesiveness (TC) (Bendersky *et al.* (2011)). Topic cohesiveness feature, $f_{TC}(D)$ is computed as:

$$f_{TC}(D) = - \sum_{w \in D} P(w|D) \log P(w|D) \quad (3.18)$$

where $p(w|D)$ is computed using a maximum likelihood estimation.

3.5 Ontology

“An ontology is an explicit specification of a conceptualization” is a prominent definition by Gruber (1995). The definition was then extended as “an ontology is explicit, formal specification of a shared conceptualization of a domain of interest” by Studer *et al.* (1998). An ontology generally consists of entities, including concepts, properties, and relations. It is the backbone to fulfill the semantic web vision (Berners-Lee *et al.* (2000); Maedche & Staab (2004)) and is a knowledge base to enable machines to communicate each other effectively. The knowledge captured in ontologies can be used to annotate data, to distinguish between homonyms and polysemy, to drive intelligent user interfaces, and even to retrieve new information. A number of ontologies are increasing day by day with new semantic web contents, because an ontology is being developed to formalize the conceptualization behind the idea of semantic web. Therefore, ontology alignment is playing an important role (Benjamins *et al.* (2002)) to achieve semantic interoperability and integration.

Chapter 4

Diversified Query Subtopic Mining

In this chapter, we present our proposed method for mining diversified query subtopics. 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 candidate subtopics, we have devised a new way of ranking based on a new novelty estimation that faithfully represents the possible search intents of the query. 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. Experimental results on NTCIR subtopic mining datasets exhibit that our proposed method outperforms the baselines, known previous methods, and the official participants of the subtopic mining tasks.

4.1 Introduction

When an information need is being formulated in a user's mind, the user issues a query as a sequence of words and submits it to the search engine. The search engine responds with a ranked list of snippet results to meet the request of users. According to user search behaviour analysis, a search query is usually

vague, ambiguous, or may entail to have multiple search intents (Song *et al.* (2009); Spärck-Jones *et al.* (2007)). Issuing the same query, different users may have different search intents, which corresponds to different subtopics (Ren *et al.* (2015)).

With an ambiguous query such as “eclipse,” users may seek different interpretations, including “eclipse IDE,” “eclipse lunar,” and “eclipse movie.” With a broad query such as “programming languages,” users may be interested in different subtopics, including “programming languages java,” “programming languages python,” and “programming languages tutorial.” However, it is not clear which subtopic of a broad query is actually desirable for a user (Wang *et al.* (2013a)). In some cases, subtopics underlying a query can be temporally ambiguous; for example, the query “US Open” is more likely to be targeting the tennis open in September and the golf tournament in June (Nguyen & Kanhabua (2014)).

Traditional information retrieval models, such as the boolean model and the vector space model, treat every input query as a clear, well-defined representation, and completely neglect any sort of ambiguities. Ignoring the users’ intents underlying a query, information retrieval models might result in top ranked documents, possibly containing too much relevant information on a particular aspect¹ of a query. As these documents cover a few subtopics or interpretations, the user may not be satisfied. In order to satisfy the user, a sensible approach is to diversify the documents considering the possible subtopics of the search query (Clarke *et al.* (2008)). The diversified retrieval models should result in a ranked list of documents that provides the maximum coverage and minimum redundancy with respect to the possible subtopics.

Identifying the subtopics underlying a query has gained much interest in recent years (Liu *et al.* (2014); Sakai *et al.* (2013a)). Several methods were proposed for mining subtopics from different resources, including the top retrieved documents, anchor texts, query logs, Wikipedia, and the related search services provided by the commercial search engines (Santos *et al.* (2010a); Wang *et al.* (2013a,c)). Query suggestions provided by commercial search engines hold some intents (Hu *et al.* (2015); Santos *et al.* (2010a)), however, suggested queries

¹Intent and Aspect are interchangeably used

are often noisy and possess a group of similar suggestions covering a single aspect of the original query. Since both query and subtopic are short in length, it is challenging to efficiently estimate the similarity between a pair of short texts and rank them accordingly. Therefore, identifying the subtopics covering possible intents underlying a query is a formidable task.

In this chapter, we address the problem of *query subtopic mining* (Sakai *et al.* (2013b)), which is defined as: “given a search query, list up its possible subtopics which specialize or disambiguate the search intents of the original query.” In our approach, we extract candidate subtopics from multiple resources and rank the subtopics covering the possible intents of the query. To estimate the relevance scores of the candidate subtopics with the query, we locally train word embedding model, extract some semantic and content-aware features followed by a supervised feature selection, and introduce a bipartite graph-based ranking (*BGR*) method. Then, we diversify the candidate subtopics with maximal marginal relevance (*MMR*) (Carbonell & Goldstein (1998)) model by balancing the relevance with the query and the novelty with other candidate subtopics. Novelty of the subtopic is estimated by combining a mutual information based similarity and categorical similarity. Experimental results on the publicly available test collections, including NTCIR-10 INTENT-2 (Sakai *et al.* (2013a)) and NTCIR-12 IMINE-2 (Yamamoto *et al.* (2016)) demonstrate that our proposed method outperforms the baselines, known previous works, and the official participants of the INTENT-2 and IMINE-2 competitions.

To summarise, our main contributions are threefold: (1) some new features based on locally trained word embedding model (in Section 4.4.2), (2) a bipartite graph-based ranking (*BGR*) method (in Section 4.4.3.2), and (3) estimating the novelty of the subtopic by combining a mutual information based similarity and categorical similarity (in Section 4.4.3.3).

The rest of this chapter is organized as follows. Section 4.2 overviews related work on query subtopic mining. Section 4.3 includes the general concepts and terminology to comprehend the readers about the contents of this chapter. We introduce our proposed diversified subtopic mining method in Sect. 4.4. Section 4.5 discusses the overall experiments and results that we obtained. Finally,

concluding remarks and some future directions of our work are described in Sect. 4.6.

4.2 Related Work

Web queries are usually short, ambiguous, and/or underspecified (Clarke *et al.* (2009); Song *et al.* (2009); Spärck-Jones *et al.* (2007)). To understand the meaning of queries, researchers define taxonomies and classify queries into predefined categories. Song *et al.* (2009) classified queries into three categories: ambiguous queries, which have more than one meaning; broad queries, which cover a variety of subtopics; and clear queries, which have a specific meaning or narrow topics.

At the query level, Broder (2002) divided query intent into navigational, informational, and transactional types. Nguyen & Kan (2007) classified queries into four general facets, including ambiguity, authority, temporal sensitivity, and spatial sensitivity. Boldi *et al.* (2008) created query-flow graph with query phrase nodes and used them for query suggestion. Query suggestion is a key technique for generating alternative queries to help users drill down to a subtopic of the original query (Mei *et al.* (2008); Zhang & Nasraoui (2006)). In contrast to query suggestion, subtopic mining focuses more on the diversity of possible subtopics of the original query rather than inferring relevant queries.

Wu *et al.* (2015) mined query subtopic from questions in the community question answering (CQA) by proposing non-negative matrix factorization (NMF) to cluster the questions and extract keywords from the cluster. Hu *et al.* (2009) leveraged the knowledge contained in Wikipedia to predict the possible subtopics for a given query. Radlinski *et al.* (2010) proposed a method for inferring query intents from query reformulations and user click-through data. Santos *et al.* (2010a) exploited the query completions of search engine to mine sub-queries (i.e. subtopics) for diversifying Web search results.

Wang *et al.* (2013a) proposed a method to mine subtopics of a query either directly from the query itself or indirectly from the top retrieved documents. In direct approach, several external resources, such as Wikipedia, open directory project (ODP), query logs, and the related search services are investigated to

mine subtopics. In indirect approach, subtopics are extracted by clustering, topic modeling, and concept-tagging of the top retrieved documents. The surrounding text of query terms in the top retrieved documents were also utilized for mining and ranking subtopics (Wang *et al.* (2013c)).

Moreno *et al.* (2014) proposed an algorithm called Dual C-Means to cluster search results in dual-representation spaces with query logs and represented the cluster label as the subtopic. Damien *et al.* (2013) proposed a method for mining and ranking the subtopic by fusing multiple resources. Kim & Lee (2015) proposed a frequent pattern-based method to mine candidate subtopics from a set of implicitly relevant documents. Our proposed method has the same premise as Moreno *et al.* (2014), Damien *et al.* (2013), and Kim & Lee (2015).

Two-level hierarchical intents based search result diversification methods were also proposed (Hu *et al.* (2015)). The method, proposed by Kim & Lee (2015) was to mine a two-level subtopic hierarchy based on hierarchical search intentions. However, our approach is to mine the flat list of the subtopic.

Neural network-based method, *Word2Vec* was proposed by Mikolov *et al.* (2013b) which represents a word in semantic space as vector is called word embedding. Words that are semantically and syntactically similar tend to be close in this embedding space. Despite the fact that previous methods leveraged many resources to mine candidate subtopics, however, their ranking methods caused some noisy and redundant subtopics in the top rank. In contrast to previous methods, we introduce locally-trained word embedding to extract semantic features and effectively diversify the candidate subtopic covering possible intents of the query by balancing the relevance and novelty.

NTCIR¹ have been organizing a research competition on *query subtopic mining* in Chinese, English and Japanese languages for the last couple of years, including NTCIR-10 INTENT-2², NTCIR-11 IMINE-1³, and NTCIR-12 IMINE-2⁴. Some approaches have been proposed by the participants exploiting multiple resources are discussed in (Kim & Lee (2013); Moreno & Dias (2013, 2016); Ullah *et al.* (2013b, 2016a); Wang *et al.* (2013b); Xue *et al.* (2013); Yue *et al.* (2016)).

¹<http://research.nii.ac.jp/ntcir/index-en.html>

²<http://research.microsoft.com/en-us/projects/intent/>

³<http://www.thuir.org/IMine/>

⁴<http://www.dl.kuis.kyoto-u.ac.jp/imine2/>

4.3 General Concepts and Terminologies

This section introduces the definitions of search intent, subtopic, and diversification.

4.3.1 Search Intent

Issuing a query into a search engine, different users seek different information needs which correspond to search intents.

Definition 4.3.1 *Search intent is defined as the intent with which a user conducts a search, meaning the information that user is searching for.*

Figure 4.1 depicts that given a query such as "apple", users may seek information related to "apple company," "apple iphone," "apple macbook," or "apple fruit." These might be the user's search intent behind the query "apple."

4.3.2 Subtopic

Subtopic is a key phrase that represents a search intent underlying a query.

Definition 4.3.2 *A set of words (i.e. phrase), which specializes or disambiguates the search intent of a search query.*

Figure 4.1 depicts that "apple iphone 6" is a subtopic representing the search intent "apple iphone" behind the query "apple." Similarly, "apple job salary" is a subtopic, which specialize the search intent "apple company."

4.3.3 Subtopic Diversification

Definition 4.3.3 *Subtopic diversification is defined as a tradeoff between finding relevant (similar) subtopics to the search query and diverse subtopics covering relevant search intent in the result set.*

In Figure 4.2, we depict the flat list of diversified subtopics covering search intents underlying the query. The challenges to generate the diversified subtopics are (a) finding relevant subtopics covering intent and (b) removing the redundant subtopics as much as possible in the result list.





Query	Search Intent	Subtopic
Apple	Apple company 	Apple pay, apple news, apple commercial, apple Walmart, apple vacations, apple job salary
	Apple iPhone 	Apple iPhone 5s, apple iPhone 6, apple iPhone unlock, apple iPhone review
	Apple MacBook 	Apple MacBook air, apple MacBook pro, apple MacBook review, apple MacBook sale
	Apple Botanical 	Apple fruit, Apple cider, Apple tree, Apple free, Oak Apple

Figure 4.1: Illustration of a query, the possible search intents, and the representative subtopics corresponding to each search intent.

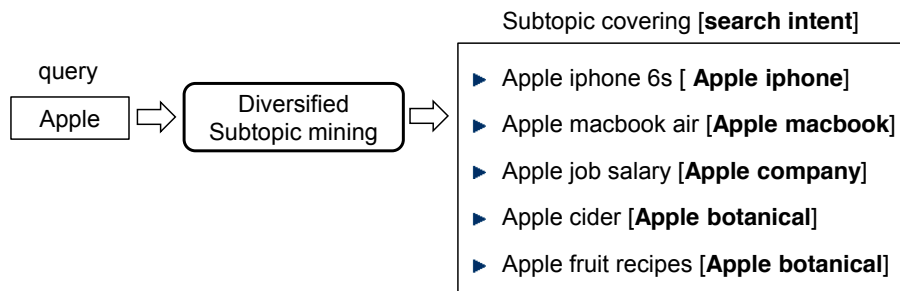


Figure 4.2: This figure demonstrates the diversified subtopics covering search intent.

4.4 Diversified Subtopic Mining

In this section, we describe our approach to subtopic mining, which is composed of candidate generation, features extraction, and ranking as depicted in Fig. 4.3. Given a query, first, we extract candidate subtopics from multiple resources. Second, to estimate the relevance of the candidate subtopics, we extract multiple semantic and content-aware features, followed by a supervised feature selection, and introduce a bipartite graph-based ranking (*BGR*) method.

Third, to cover the possible search intents of the query, we produce a diversified ranked list of subtopics by balancing the relevance and the novelty. We propose to estimate the novelty of subtopic by combining a mutual information based similarity through the *Jensen-Shannon divergence* tuned for short texts through the probability distributions of terms in the top retrieved documents from a search engine and cluster-based categorical similarity. Our detail method is articulated as follows:

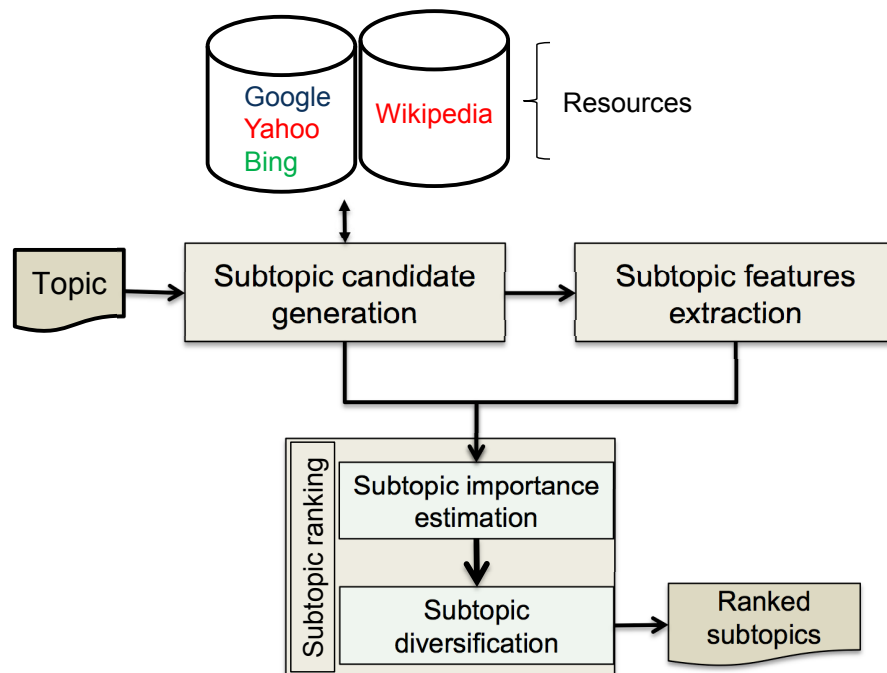


Figure 4.3: Diversified subtopic mining flow

4.4.1 Subtopic Candidate Generation

Inspired by the work of Santos et al. (Santos *et al.* (2010a)), we hypothesize that *suggested queries* in across search engines hold some intents of the query. Given a query, we collect all the suggested queries from search engines. If a query is matched with the title of a Wikipedia disambiguation page¹, we also extract the different meanings from that page. Then, we aggregate them by

¹[https://en.wikipedia.org/wiki/Apple_\(disambiguation\)](https://en.wikipedia.org/wiki/Apple_(disambiguation))

filtering out the duplicates or wrongly represented ones, and consider them as candidate subtopics. To filter out the duplicates or wrongly represented ones, we apply canonicalization on the suggested queries using Krovetz (Krovetz (1993)) stemmer and remove those ones which are part of the query or exactly similar with the query. For example, given a query “old coins,” we generate a list of candidate subtopics including “old coins sell,” “old gold coins,” “old coins for sale,” and “old coins prices.”

4.4.2 Subtopic Features Extraction

Let $q \in \mathcal{Q}$ represents a query and $\mathcal{S} = \{s_1, s_2, \dots, s_k\}$ represents a set of candidate subtopics generated in Sect. 4.4.1. Both query and candidate subtopics are short in length and it is a challenging task to estimate the similarity between a pair of short texts. We extract multiple local and global features, which are broadly organized as word embedding and content-aware features. While the content-aware features are standard features commonly used in the literature for learning to rank for Web search (Liu (2009)), the word embedding based features are specifically proposed to estimate the relevance of candidate subtopics for a query.

We propose three semantic features based on locally-trained word embedding and make use of *word2vec*¹ model (Mikolov *et al.* (2013b)). In order to capture the importance of the semantic matching of a query with a subtopic, we first propose a new feature, the maximum word similarity (*MWS*) as follows:

$$f_{MWS}(q, s) = \frac{1}{|q|} \sum_{t \in q} sem(t, s) \quad (4.1)$$

where

$$sem(t, s) = \max_{w \in s} f_{sem}(\vec{t}, \vec{w})$$

¹*word2vec* (<https://code.google.com/p/word2vec/>)

where \vec{t} and \vec{w} are the word vector representations from *word2vec* model, corresponding to two words t and w , respectively. The function f_{sem} returns the cosine similarity between two word vectors.

To measure the global importance of a query with a subtopic, we propose our second feature, the mean vector similarity (*MVS*) as follows:

$$f_{MVS}(q, s) = f_{sem}\left(\frac{1}{|q|} \sum_{t \in q} \vec{t}, \frac{1}{|s|} \sum_{w \in s} \vec{w}\right) \quad (4.2)$$

Our third proposed feature, the uncommon word similarity (*UWS*) is defined through the similarity of the uncommon word of query and subtopic as follows:

$$f_{UWS}(q, s) = f_{sem}\left(\frac{1}{|q_u|} \sum_{t \in q_u} \vec{t}, \frac{1}{|s_u|} \sum_{w \in s_u} \vec{w}\right) \quad (4.3)$$

where query q and subtopic s represent two sets of words, respectively. Then, q_u is defined as $q - (q \cap s)$ and s_u is defined as $s - (q \cap s)$. These three semantic features in Eqs. (4.1), (4.2), and (4.3) are complementary to each others.

Among content-aware features, we extract term frequency, language modeling, term dependency, lexical, and Web hit-count based features. Term frequency (*TF*) based features are directly computed by scoring the occurrences of the terms of query q in a subtopic s . Term frequency based features include *DPH* (Amati (2003)), *PL2* (Amati (2003)), and *BM25* (Robertson & Zaragoza (2009)). Language modeling (*LM*) based features include Kullback-Leibler (*KL*) (Lafferty & Zhai (2001)), query likelihood with Jelinek-Mercer (*QLM-JM*) (Zhai & Lafferty (2001)), subtopic likelihood with Jelinek-Mercer (*SLM-JM*) (Zhai & Lafferty (2001)), query likelihood with Dirichlet smoothing (*QLM-DS*) (Zhai & Lafferty (2001)), and subtopic likelihood with Dirichlet smoothing (*SLM-DS*) (Zhai & Lafferty (2001)). Term dependency (*TD*) based features include term-dependency Markov random field (*MRF*) (Metzler & Croft (2005)) and Tri-gram dependency.

To measure the lexical similarity (*LS*) between a query and a subtopic, we extract features based on edit distance (*EDS*), sub-string match (*SSM*) (Metzler &

4.4 Diversified Subtopic Mining

Kanungo (2008)), term overlap (*TO*) (Metzler & Kanungo (2008)), term synonym overlap (*TSO*) (Metzler & Kanungo (2008)), vector space model (*VSM*), and coordinate level matching (*CLM*) (Salton & Buckley (1988)).

If a subtopic is frequently mentioned in the Web pages, then that subtopic might be important than others. According to this intuition, we make use of search engine hit count (*HC*) to estimate features including normalized hit count (*NHC*), point-wise mutual information (*PMI*), and word co-occurrence (*WC*). To encode a prior knowledge (*PK*) about individual subtopic, we also extract simple query independent features including voting, reciprocal rank (*RR*), average term length (*ATL*), topic cohesiveness (*TC*) (Bendersky *et al.* (2011)), and subtopic length (*SL*). For each query-subtopic pair, 27 features are extracted based on word embedding and content-aware relevance as stated in Table 4.1.

Table 4.1: Word embedding and content-aware based features in this work.

Features	Type	Total
MWS, MVS, UWS	<i>Word2Vec</i>	3
DPH, PL2, BM25	<i>TF</i>	3
KL, QLM-JM, SLM-JM, QLM-DS, SLM-DS	<i>LM</i>	5
MRF, Tri-Gram	<i>TD</i>	2
EDS, SSM, TO, TSO, VSM, CLM	<i>LS</i>	6
NHC, PMI, WC	<i>HC</i>	3
Voting, RR, ATL, TC, SL	<i>PK</i>	5
		27

4.4.3 Subtopic Ranking

For a pair of query q and candidate subtopic s , we extract all the features described in Sect. 4.4.2 and represent those in a feature vector, $\mathcal{F}_{q,s} = \{f_{DPH}(q, s), f_{PL2}(q, s), \dots, f_{UWS}(q, s)\}$. Therefore, for a query q , we have a feature matrix, $\mathcal{MF} = \{\mathcal{F}_{s_1}, \mathcal{F}_{s_2}, \dots, \mathcal{F}_{s_k}\}$, corresponding to a set of candidate subtopics, $\mathcal{S} = \{s_1, s_2, \dots, s_k\}$. We normalize each feature vector using *MinMax* normalization technique. To estimate the diversified rank of a subtopic, first, we employ a supervised feature selection method to remove noisy and redundant features. Second, we introduce a bipartite graph-based ranking approach for estimating

the relevance of candidate subtopic. Finally, we make use of *MMR* model to produce a diversified ranked list of subtopics covering the possible intents of the query by balancing the relevance of the candidate subtopic with the query and novelty with other subtopics.

4.4.3.1 Supervised Feature Selection

Supervised feature selection is an important technique to determine the best set of features by reducing the noisy, redundant or highly correlated features in a large feature set. We make use of elastic-net regularized regression method due to its better performance over Lasso and Ridge regression (Zou & Hastie (2005)). Given a parameter α strictly between 0 and 1, and a nonnegative λ , elastic-net solves the following optimization problem:

$$\min_{\beta_0, \beta} \left(\frac{1}{2M} \sum_{i=1}^M (y_i - \beta_0 - \mathcal{F}_i^T \beta)^2 + \lambda \sum_{j=1}^p \left(\frac{1-\alpha}{2} \beta_j^2 + \alpha \|\beta_j\| \right) \right) \quad (4.4)$$

where M is the number of samples, \mathcal{F}_i^T is the transpose of feature vector of the i -th sample, and $y_i \in \{0, 1\}$ is the label of the i -th sample. In our case, each sample is a query-subtopic pair. We train elastic-net on query-subtopic pairs' feature vectors and choose those features whose coefficients β are positive.

4.4.3.2 Subtopic Relevance Estimation

To estimate the relevance of the candidate subtopics for a query, we introduce a bipartite graph-based ranking (*BGR*) approach with considering the features selected by elastic-net in Sec. 4.4.3.1.

Bipartite Graph based Ranking Many real applications can be modeled as a bipartite graph, including Video shots and Tags (YANAI *et al.* (2015)), Queries and URLs in query logs, Entities and Co-List (Cao *et al.* (2011)) in a Web page, Phenotypes and Diseases (Ullah *et al.* (2015)) in bioinformatics.

We hypothesize that a relevant subtopic should be ranked at the higher position by multiple effective features, and intuitively, an effective feature

should be weighted higher by multiple relevant subtopics. Large weight should be given to a subtopic that tends to be ranked highly by a group of effective features, and vice versa. Therefore, there is a weight propagation of features to subtopics and subtopics to features. On these intuitions, we represent a set of features and a set of candidate subtopics as a bipartite graph and introduce weight propagation from both sides of the bipartite graph. Given a set of features and a set of candidate subtopics, we propose a bipartite graph-based ranking (BGR) method to estimate the global importance of candidate subtopics by aggregating the local importance of the individual feature.

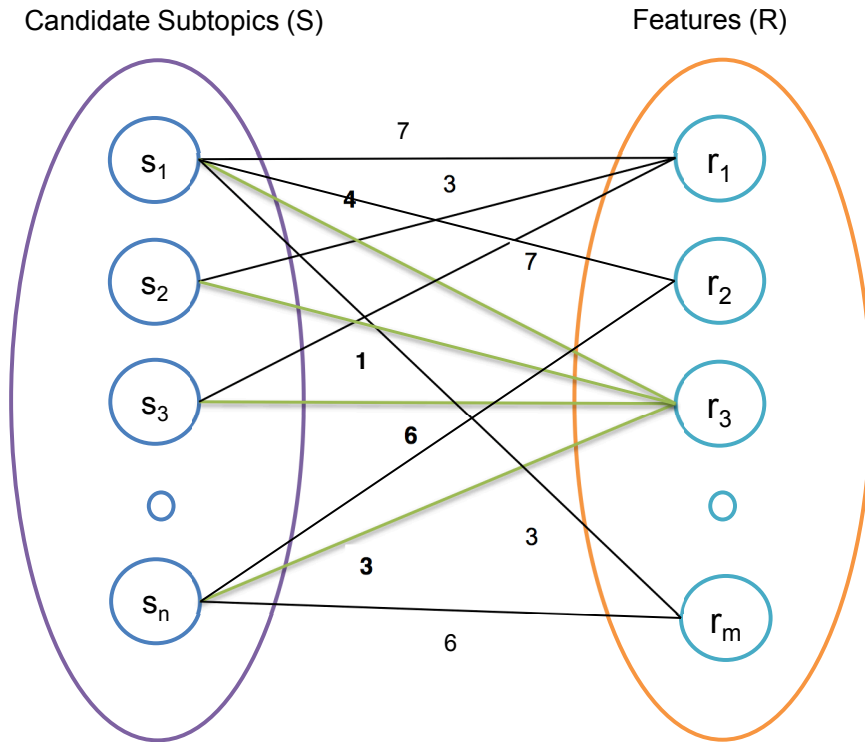


Figure 4.4: Bipartite graph based representation of subtopics and features

Let $\mathcal{R} = \{r_1, r_2, \dots, r_m\}$ be a set of features and $\mathcal{S} = \{s_1, s_2, \dots, s_n\}$ be a set of candidate subtopics. Consider $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is a bipartite graph as depicted in Figure 4.4, where vertex set \mathcal{V} is composed of two disjoint sets \mathcal{R} and \mathcal{S} , such that each edge in \mathcal{E} connects a vertex in \mathcal{R} to a vertex in \mathcal{S} ; that is, there is no edge between two vertices in the same set. For each feature $r_i \in \mathcal{R}$, we obtain a

ranked list of subtopics L_{r_i} . The weight w_{ij} of the edge between two vertices $r_i \in \mathcal{R}$ and $s_j \in \mathcal{S}$ is defined as follows:

$$w_{ij} = \frac{1}{\sqrt{\log_2(\text{rank}(L_{r_i}, s_j) + 2.0)}} \quad (4.5)$$

where the function, $\text{rank}(L_{r_i}, s_j)$ returns the position of the subtopic s_j in the ranked list L_{r_i} for the feature r_i . The reciprocal rank of the subtopic is considered to assign high importance to the subtopic in the higher rank position.

The bipartite graph \mathcal{G} is represented as a bi-adjacency $m \times n$ matrix, \mathcal{M} . We assign an initial weight to each vertex in the bipartite graph \mathcal{G} . Therefore, the weight vector corresponding to the vertices in the set of features \mathcal{R} is R , which is initialized as uniform values (i.e. $1/m$ where m is the number of features). The weight vector corresponding to the vertices in the set of subtopics \mathcal{S} is S , which is initialized by applying, either the equations (4.1) or (4.2) between a query and a candidate subtopic.

For a bipartite graph, there is a natural random walk on the graph with the transition probability from both sides (Deng *et al.* (2009b)). The transition matrix from \mathcal{R} to \mathcal{S} is defined as $\mathcal{W}_1 = \mathcal{D}_{\mathcal{R}}^{-1}\mathcal{M}$, where $\mathcal{D}_{\mathcal{R}}$ is the diagonal matrix with its (i, i) -element equal to the sum of the i -th row of \mathcal{M} . Similarly, the transition matrix from \mathcal{S} to \mathcal{R} is defined as $\mathcal{W}_2 = \mathcal{D}_{\mathcal{S}}^{-1}\mathcal{M}^T$, where $\mathcal{D}_{\mathcal{S}}$ is the diagonal matrix with its (j, j) -element equal to the sum of the j -th column of \mathcal{M} .

The weight propagation from the set of candidate subtopics \mathcal{S} to the set of features \mathcal{R} is represented as follows:

$$R_{k+1} = \lambda_1 \mathcal{W}_1 S_k + (1 - \lambda_1) R_0 \quad (4.6)$$

where $0 < \lambda_1 < 1$ is a parameter, which is used to combine the initial and the propagated scores of the features. R_0 is the vector of the initial scores of the features, S_k denotes the vector of the subtopics score after k -th iterations, and R_{k+1} denotes the vector of the features score after $(k + 1)$ -th iterations.

Similarly, the weight propagation from the set of features to the set of

candidate subtopics is represented as follows:

$$S_{k+1} = \lambda_2 \mathcal{W}_2 R_{k+1} + (1 - \lambda_2) S_0 \quad (4.7)$$

where $0 < \lambda_2 < 1$ is a parameter, which is used to combine the initial and the propagated scores of the candidate subtopics. S_0 is the vector of the initial scores of the subtopics and S_{k+1} denotes the vector of the subtopics score after $(k + 1)$ -th iterations.

Considering the vertices in \mathcal{S} , by substituting equation. (4.6) for R_{k+1} in equation. (4.7), the weight equation of subtopics is represented as follows:

$$\begin{aligned} S_{k+1} &= \lambda_2 \mathcal{W}_2 [\lambda_1 \mathcal{W}_1 S_k + (1 - \lambda_1) R_0] + (1 - \lambda_2) S_0 \\ &= \lambda_1 \lambda_2 \mathcal{W}_2 \mathcal{W}_1 S_k + \lambda_2 (1 - \lambda_1) \mathcal{W}_2 R_0 + (1 - \lambda_2) S_0 \end{aligned} \quad (4.8)$$

The closed form of equation. (4.8) is defined as follows:

$$\begin{aligned} S_{k+1} &= (\lambda_1 \lambda_2 \mathcal{W}_2 \mathcal{W}_1)^{k+1} S_0 + \sum_{t=0}^k (\lambda_1 \lambda_2 \mathcal{W}_2 \mathcal{W}_1)^t \\ &\quad [\lambda_2 (1 - \lambda_1) \mathcal{W}_2 R_0 + (1 - \lambda_2) S_0] \end{aligned} \quad (4.9)$$

Since Eigen values of the stochastic matrices \mathcal{W}_1 and \mathcal{W}_2 are in $[-1, 1]$, the equation. (4.9) is converged. Therefore,

$$\begin{aligned} \lim_{k \rightarrow \infty} (\lambda_1 \lambda_2 \mathcal{W}_2 \mathcal{W}_1)^{k+1} &= 0 \\ \lim_{k \rightarrow \infty} \sum_{t=0}^k (\lambda_1 \lambda_2 \mathcal{W}_2 \mathcal{W}_1)^t &= (I - \lambda_1 \lambda_2 \mathcal{W}_2 \mathcal{W}_1)^{-1} \end{aligned} \quad (4.10)$$

When $k \rightarrow \infty$ and $S_{k+1} \rightarrow S^*$, we have

$$S^* = (I - \lambda_1 \lambda_2 \mathcal{W}_2 \mathcal{W}_1)^{-1} [(1 - \lambda_1) \lambda_2 \mathcal{W}_2 R_0 + (1 - \lambda_2) S_0] \quad (4.11)$$

Given λ_1 , λ_2 , \mathcal{W}_1 , \mathcal{W}_2 , R_0 , and S_0 , we estimate the scores S^* directly by applying equation. (4.11). These scores S^* are considered as the relevance scores, $\text{rel}(q, \mathcal{S})$ of a set of candidate subtopics \mathcal{S} with the query q and eventually utilized in subtopic diversification.

4.4.3.3 Subtopic Diversification

To produce a diversified ranked list of subtopics by balancing the relevance and novelty, we make use of *MMR* model. The *MMR* regards the ranking problem as a procedure of successively selecting the “best” unranked subtopic. When searching for the next best subtopic, the *MMR* model chooses not the most relevant one, however, the one that best balances the relevance and novelty. Novelty means that a subtopic is new compared to those already selected and ranked.

Given a relevance function $rel(\cdot, \cdot)$ and a novelty function $novelty(\cdot, \cdot)$, the *MMR* model can be defined as follows:

$$s_i^* = \operatorname{argmax}_{s_i \in D \setminus C_i} \gamma rel(q, s_i) + (1 - \gamma) novelty(s_i, C_i) \quad (4.12)$$

where γ is a combining parameter and $\gamma \in [0, 1]$. D is the relevance oriented ranked list of subtopics retrieved by equation. (4.11). C_i is the collection of subtopics that have already been selected at the i -th iteration and initially empty. Then,

$$C_{i+1} = C_i \cup \{s_i^*\}$$

where s_i^* is the subtopic ranked at the i -th position. The function, $novelty(s_i, C_i)$ tries to measure the novelty of the subtopic s_i given the collection C_i .

We find the maximum similarity value for subtopic s with all the selected subtopics $s' \in C_i$, and flip the sign as the novelty score as follows:

$$novelty(s_i, C_i) = - \max_{s' \in C_i} sim(s_i, s') \quad (4.13)$$

Since both subtopics s and s' are short in length and they might not be lexically similar. We hypothesize that if two subtopics represent the similar meaning, even though they are not lexically similar, they may belong to the similar categories and retrieve similar kinds of documents from a search engine.

Mutual information between two probability distributions of words may represent the contextual similarity between two subtopics. Therefore, we propose to estimate the novelty of a subtopic by combining the contextual similarity and categorical similarities as follow:

$$novelty(s_i, C_i) = -\max_{s' \in C_i} \left(1.0 - \sqrt{JSD(s_i, s')} + \sum_{x \in X} \frac{[(s_i, s') \in x]}{|x|} \right) \quad (4.14)$$

where $JSD(s_i, s')$ is estimated through the *Jensen-Shannon divergence* of the word probability distributions of the top-k documents refer to the subtopics s_i and s' . X is the set of clusters obtained by applying the frequent phrase-based soft clustering (Osiński *et al.* (2004)) on the set of candidate subtopics \mathcal{S} , $|x|$ is the number of subtopics belong to the cluster x , and $[(s_i, s') \in x] = 1$ if true, zero, otherwise.

$JSD(s, s')$ is defined as follows:

$$\begin{aligned} JSD(s, s') &= JSD(P, Q) \\ &= \frac{1}{2} \sum_{i=1}^{|V|} \left(P(i) \log \frac{P(i)}{T(i)} + Q(i) \log \frac{Q(i)}{T(i)} \right) \end{aligned} \quad (4.15)$$

where $T = \frac{1}{2}(P + Q)$. P and Q refer the word probability distributions, extracted from the top- K retrieved documents from the search engine for subtopics s and s' . V is the set of words in the vocabulary, collected from the titles and snippets of the documents corresponding to two subtopics s and s' . We choose *Jensen-Shannon divergence* over *Kullback-Leibler divergence*, because of its symmetric similarity estimate.

4.5 Experiments and Discussion

In this section, we evaluate our proposed method with different settings on the NTCIR-10 INTENT-2 (Sakai *et al.* (2013b)) and NTCIR-12 IMINE-2 (Yamamoto *et al.* (2016)) English Subtopic Mining test collections and compare the performance with the previous works.

4.5.1 Dataset

The NTCIR-10 INTENT-2 and NTCIR-12 IMINE-2 English Subtopic Mining test collections include a set of 50 and 100 topics (i.e. queries), respectively. Each topic is labelled by a set of intents with probabilities and for each intent, there is a set of subtopics as relevance judgement. The statistics of the topics, intents, and subtopics of the INTENT-2 and IMINE-2 datasets are stated in Table 4.2. For example, a topic “grilling,” which is labelled by intents with probabilities and a set of subtopics under each intent depicted in Figure 4.5. It shows that topic “grilling” has several intents with probabilities and there is a set of subtopics as examples for each intent.

Table 4.2: Statistics of Topics, Intents, and Subtopics of NTCIR-10 INTENT-2 and NTCIR-12 IMINE-2 Datasets

	INTENT-2	IMINE-2
Topics	50	100
Intents	392	533
Subtopics	5,410	1652

```

<topic number="0410">
  <query>grilling</query>
  <intent number="1" probability="0.175000">
    <description>grilling recipes</description>
    <examples>summer grilling recipes,...</examples>
  </intent>
  <intent number="2" probability="0.165000">
    <description> grilling barbecue</description>
    <examples>meat for grilling,...</examples>
  </intent>
</topic>

```

Figure 4.5: A topic “grilling” from INTENT-2 dataset, which is labelled by its intents with probabilities and a set of subtopics under each intent

Both INTENT-2 and IMINE-2 organizers collected query suggestions and query completions from *Google*, *Yahoo*, and *Bing* search engines corresponding to the set of topics and included in the datasets as resources for fairly comparing

the methods using these datasets. We made use of the query suggestions that were included in INTENT-2 and IMINE-2 datasets.

To estimate some global features, including inverse document frequency (*IDF*) and corpus frequency (*CF*), we indexed the *clueweb12-b13* (Callan *et al.* (2009)) corpus employing Indri Search Engine (Strohman *et al.* (2005)) and utilized accordingly. To estimate the features, including Eqs. (4.1), (4.2), and (4.3), for each topic, we retrieved the top-1000 documents from the *clueweb12-b13* corpus based on language model and locally trained word embedding based on *word2vec*. The parameters in the *word2vec* tool were Skip-gram architecture, window width of 10, dimensionality of 200, and the sampling threshold of 10^{-3} . For estimating the novelty function in equation. (4.14), we queried the *Bing Search API*¹ and collected the top-50 documents corresponding to all the topics and the candidate subtopics.

4.5.2 Evaluation Metrics

We evaluated our proposed method by estimating I-rec, D-nDCG, and D#-nDCG at the cutoff rank 10. I-rec@10 measures the diversity of the returned subtopics, which shows how many percentages of intents can be found. D-nDCG@10 measures the overall relevance across all intents considering the subtopic ranking. D#-nDCG@10 is a combination of I-rec@10 (50%) and D-nDCG@10 (50%). It is used as the primary evaluation metric by the INTENT-2 and IMINE-2 task organizers (Sakai *et al.* (2013a); Yamamoto *et al.* (2016)). The advantages of D#-nDCG@10 over other diversity metrics (e.g. a-nDCG and Intent-Aware metrics) are discussed in (Sakai *et al.* (2013a)). In our experiments, we utilized NTCIREVAL (Sakai (2011)), the tool provided by the NTCIR organizers to compute the above three metrics, in which the default setting was used. Moreover, we made use of two-tailed paired t-test for statistical significance analysis, where the significance level is 0.05 (Sakai (2014)).

¹<https://datamarket.azure.com/dataset/bing/search>

4.5.3 Experimental Settings

We designed three experiments to evaluate the usefulness of our proposed method. In experiment 1, we discriminatively evaluated our proposed method by highlighting three contributions on INTENT-2 and IMINE-2 datasets in different settings, including *Baseline*, *W2V-LRM-Cosine*, *W2V-BGR-Cosine*, and *W2V-BGR-JSD*. *Baseline* included standard content-aware features, linear ranking method (*LRM*), and cosine similarity-based novelty function. *W2V-LRM-Cosine* extended *Baseline* by including word embedding based features. *W2V-BGR-Cosine* included a bipartite graph-based ranking (*BGR*) method in place of *LRM* in *W2V-LRM-Cosine*. Our proposed *W2V-BGR-JSD* included novelty equation. 4.14 in place of cosine similarity-based novelty function in *W2V-BGR-Cosine*. Moreover, to show the effectiveness of the feature selection, we also evaluated our proposed method with or without considering feature selection.

In experiment 2, we compared the performance of our proposed method *W2V-BGR-JSD* on INTENT-2 dataset with the known related methods, including Kim & Lee (2015), Moreno *et al.* (2014), and Damien *et al.* (2013), and the baselines, including query completions (BingC, GoogleC and YahooC), query suggestion (BingS), and a simple merging strategy (MergeBGY).

In experiment 3, we compared the performance of our proposed method *W2V-BGR-JSD* with the official participants of INTENT-2 (Sakai *et al.* (2013a)) and IMINE-2 competitions (Yamamoto *et al.* (2016)).

4.5.4 Important Features and Parameter Tuning

To select the important features, we prepared the training samples by choosing 10 queries, including the corresponding subtopics from the relevance judgement of INTENT-2 and IMINE-2 datasets, respectively. We extracted in total 27 features and employed elastic-net for feature selection. The selected features for INTENT-2 dataset were as follows: *MWS*, *MVS*, *UWS*, *DPH*, *QLM-JM*, *SLM-DS*, *MRF*, *SSM*, *TSO*, *NHC*, *WC*, *RR*, and *ATL*. Similarly, the selected features for IMINE-2 dataset were as follows: *MWS*, *MVS*, *UWS*, *DPH*, *BM25*, *MRF*, *TO*, *TSO*, *NHC*, *WC*, *Voting*, and *TC*. It turns out that our proposed features *MWS*,

MVS, and *UWS* were selected for both of the datasets and are important in subtopic ranking.

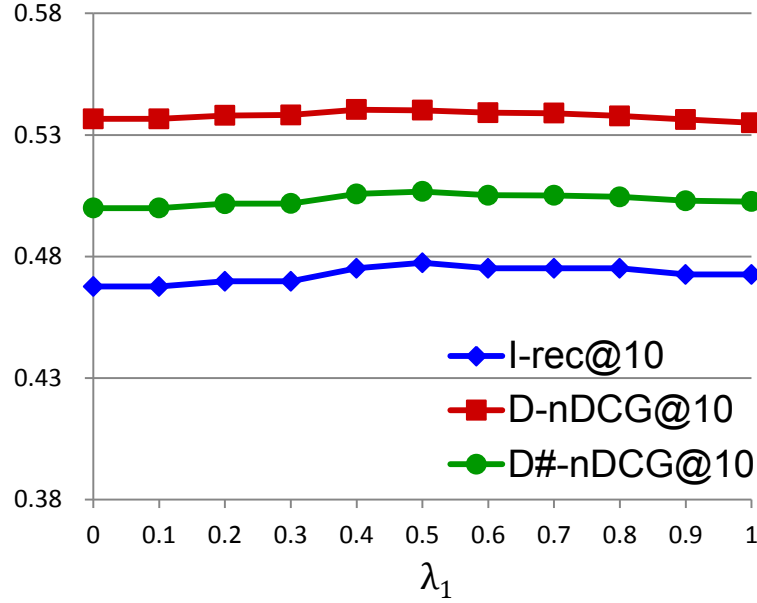


Figure 4.6: An empirical analysis of the parameter λ_1 in equation. (4.6) for our proposed method bipartite graph based *BGR* ranking on INTENT-2 dataset. The X-axis indicates the parameter λ_1 and the Y-axis indicates the corresponding scores of I-rec, D-nDCG, and D#-nDCG at the cutoff rank 10.

There were some optimization parameters in our proposed method, namely λ_1 , λ_2 , and γ in Eqs. (4.6), (4.7), and (4.12), respectively. With empirical evaluation, we found the optimal values of these parameters for both INTENT-2 and IMINE-2 datasets. To find out the optimal values of λ_1 , λ_2 and γ , we fixed the λ_2 to 0.5 and changed the value of λ_1 at a rate of 0.1 from 0 to 1. The evaluation result is depicted in Figure 4.6. It turns out that the curve of diversity measure I-rec@10 is smooth from 0.6 to 0.8 values of λ_1 . Therefore, the optimal insensitive value of λ_1 is 0.8, which reflects a high importance of the propagated weights of features than the initial weights in equation. (4.6).

Then, we fixed the λ_1 to 0.8 and changed the value of λ_2 at a rate of 0.1 from 0 to 1. The evaluation result is depicted in Figure 4.7. It demonstrates that the curve of diversity measure I-rec@10 is smooth from 0.3 to 0.5 for values of λ_2 . Therefore, the optimal value of λ_2 is 0.4, which reveals a higher

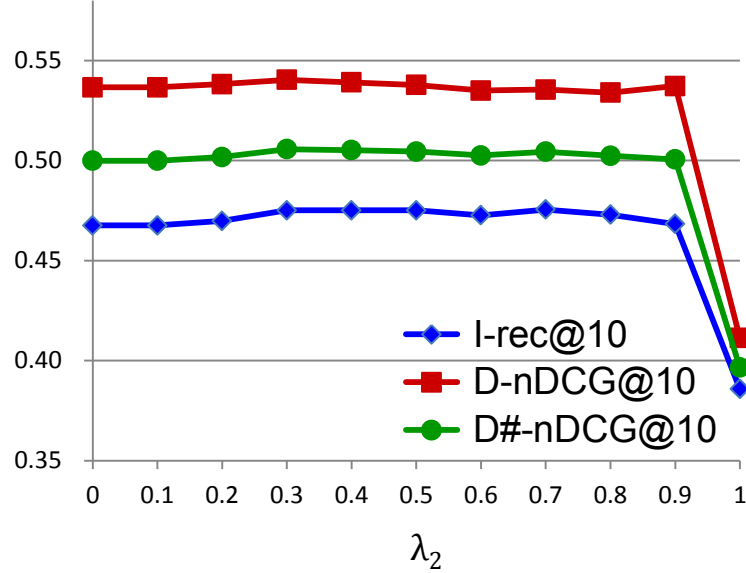


Figure 4.7: An empirical study of the parameter λ_2 in equation. (4.7). The X-axis indicates the parameter λ_2 and the Y-axis indicates the corresponding scores of I-rec, D-nDCG, and D#-nDCG at the cutoff rank 10.

importance of the initial weight than propagated weight for candidate subtopics in equation. (4.7).

In equation. (4.12), the diversification parameter γ balances the relevance and novelty of candidate subtopics. With the optimal values of λ_1 to 0.8 and λ_2 to 0.4, we changed the values of γ at a rate of 0.05 from 0.5 to 1. The evaluation result is depicted in Figure 4.8. It shows that if we increase the value of γ , both diversity measures I-rec@10 and relevance measure D-nDCG@10 increase. However, around a value of 0.80 of γ , I-rec@10 decreases and D-nDCG@10 increases. We found the highest value of D#-nDCG@10 metric at 0.85 for the parameter γ , which reflected that *MMR* model assigned high scores to relevance than novelty for subtopics diversification. Therefore, the optimal values of the parameters λ_1 , λ_2 , and γ for INTENT-2 dataset are 0.80, 0.40, and 0.85, respectively. Similarly, we found the optimal values of the parameters λ_1 , λ_2 , and γ for IMINE-2 dataset are 0.60, 0.70, and 0.80, respectively.

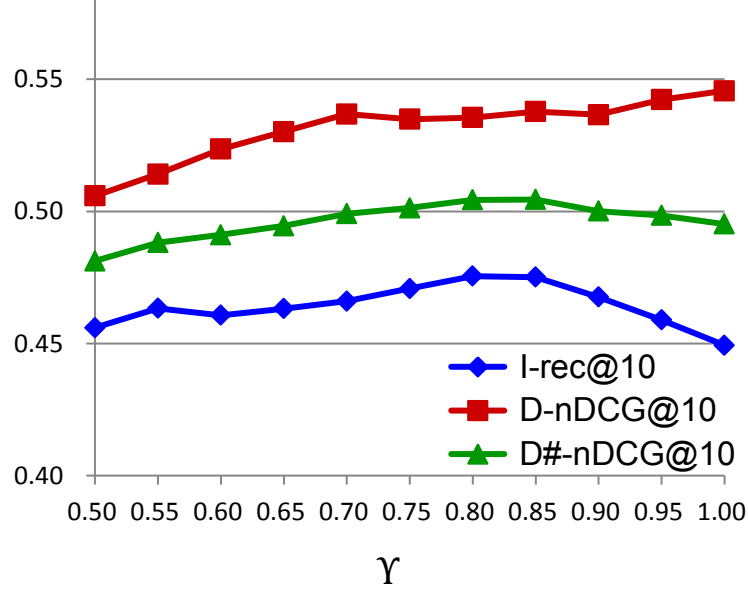


Figure 4.8: Sensitivity analysis of the diversification parameter γ . The X-axis indicates the different values of the parameter γ and the Y-axis indicates the corresponding scores of I-rec, D-nDCG, and D#-nDCG at the cutoff rank 10.

4.5.5 Experimental Results

To fairly compare with the baselines and previous works, we evaluated our proposed method on INTENT-2 topics by utilizing the parameters learned for IMINE-2 topics. Similarly, we evaluated our proposed method on IMINE-2 topics by utilizing the parameters learned for INTENT-2 topics. In addition, we excluded the topics which were used for feature selection.

Experiment 1: we evaluated our proposed method in different settings, including *Baseline*, *W2V-LRM-Cosine*, *W2V-BGR-Cosine*, and *W2V-BGR-NOV* on INTENT-2 and IMINE-2 datasets. For INTENT-2 dataset, the comparative performances are reported in Table 4.3. It turns out that with two-tailed paired t-test ($p < 0.05$), *W2V-BGR-JSD* significantly outperforms *W2V-BGR-Cosine*, *W2V-LRM-Cosine*, and *Baseline* in terms of D-nDCG@10 and D#-nDCG@10, however, *W2V-BGR-Cosine* and *W2V-LRM-Cosine* are indistinguishable from *W2V-BGR-JSD* in terms of I-rec@10.

For IMINE-2 dataset, the comparative performances are shown in Table 4.4. With two-tailed paired t-test ($p < 0.05$), it demonstrates that *W2V-BGR-JSD* signif-

4.5 Experiments and Discussion

Table 4.3: Performance comparison of our proposed method in different settings on NTCIR-10 INTENT-2 dataset at the cutoff rank 10. The best result is in bold. † indicates statistically significant and \diamond indicates statistically indistinguishable from the best.

Method	I-rec	D-nDCG	D#-nDCG
Our proposed (<i>W2V-BGR-JSD</i>)	0.4745 [†]	0.5394 [†]	0.5048 [†]
<i>W2V-BGR-Cosine</i>	0.4616 \diamond	0.5089	0.4850
<i>W2V-LRM-Cosine</i>	0.4466 \diamond	0.4780	0.4626
<i>Baseline</i>	0.4152	0.4555	0.4360

Table 4.4: Performance comparison of our proposed method in different settings on NTCIR-12 IMINE-2 dataset at the cutoff rank 10. The best result is in bold. † indicates statistically significant.

Method	I-rec	D-nDCG	D#-nDCG
Our Proposed (<i>W2V-BGR-JSD</i>)	0.8235 [†]	0.6911 [†]	0.7581 [†]
<i>W2V-BGR-Cosine</i>	0.7319	0.6312	0.6837
<i>W2V-LRM-Cosine</i>	0.7216	0.6217	0.6704
<i>Baseline</i>	0.6928	0.5762	0.6377

icantly outperforms *W2V-BGR-Cosine*, *W2V-LRM-Cosine*, and *Baseline* in terms of I-rec@10, D-nDCG@10, and D#-nDCG@10 metrics.

To show the effectiveness of the feature selection, we described the comparative performances of our proposed method with or without employing feature selection on INTENT-2 and IMINE-2 datasets in Table 4.5. It reveals that *W2V-BGR-JSD* significantly outperforms its variants without including feature selection for all metrics on IMINE-2 datasets. However, *W2V-BGR-JSD* is statistically indistinguishable from its variants for D-nDCG@10 and D#-nDCG@10 on INTENT-2 dataset.

Experiment 2: we compared our proposed method (*W2V-BGR-JSD*) with the known related methods, including Kim & Lee (2015), Moreno *et al.* (2014), Damien *et al.* (2013), and the baselines including BingC, GoogleC, YahooC, BingS, and MergeBGY for INTENT-2 dataset. The comparative performances are shown in Table 4.6. Overall, *W2V-BGR-JSD* outperforms the baselines, Kim & Lee (2015), Moreno *et al.* (2014), and Damien *et al.* (2013) in terms of I-rec@10, D-nDCG@10, and D#-nDCG@10 metrics.

4.5 Experiments and Discussion

Table 4.5: Performance comparison of our proposed method with/without feature selection (FS) on NTCIR-10 INTENT-2 and NTCIR-12 IMINE-2 datasets at the cutoff rank 10. The best result is in bold. † indicates statistically significant and ◊ indicates statistically indistinguishable from the best.

	Method	I-rec	D-nDCG	D#-nDCG
NTCIR-10 INTENT-2	W2V-BGR-JSD	0.4745 [†]	0.5394	0.5048
	Without FS	0.4390	0.5279 [◊]	0.4794 [◊]
NTCIR-12 IMINE-2	W2V-BGR-JSD	0.8235 [†]	0.6911 [†]	0.7581 [†]
	Without FS	0.7305	0.6245	0.6801

Table 4.6: Comparative performance of our proposed method with baselines and known previous methods on NTCIR-10 INTENT-2 dataset at the cutoff rank 10. The best result is in bold. † indicates statistically significant and ◊ indicates statistically indistinguishable from the best.

	Method	I-rec	D-nDCG	D#-nDCG
Our Proposed	W2V-BGR-JSD	0.4745 [†]	0.5394 [†]	0.5048 [†]
Baselines	BingS	0.3068	0.2787	0.2928
	BingC	0.3231	0.3268	0.3250
	GoogleC	0.3735	0.3841	0.3788
	YahooC	0.3829	0.3815	0.3822
	MergeBGY	0.3365	0.3181	0.3273
Previous Methods	Kim & Lee (2015)	0.4032	0.3681	0.3788
	Moreno <i>et al.</i> (2014)	0.4249	0.4221	0.4225
	Damien <i>et al.</i> (2013)	0.4587 [◊]	0.3625	0.4106

With two-tailed paired t-test ($p < 0.05$), in terms of diversity (i.e. I-rec@10), *W2V-BGR-JSD* significantly outperforms the baselines, Kim & Lee (2015), and Moreno *et al.* (2014), however, Damien *et al.* (2013) is statistically indistinguishable from *W2V-BGR-JSD*. Despite the fact that Kim & Lee (2015) proposed simple rules and popularity-based ranking, Moreno *et al.* (2014) applied Dual-C means clustering, and Damien *et al.* (2013) employed *Jaccard similarity* based hierarchical clustering to mine candidate subtopics, which often cause irrelevant and redundant candidates, however, our proposed novelty function in equation. (4.14), which is derived by combining *Jensen-Shannon divergence* tuned for short texts eliminates redundant candidate subtopics and benefits for diverse relevant subtopics at the top rank.

In terms of relevance (i.e. D-nDCG@10), *W2V-BGR-JSD* shows statistically significant improvement over the baselines, Kim & Lee (2015), Moreno *et al.* (2014), and Damien *et al.* (2013). To estimate the relevance of the candidate subtopics, our proposed bipartite graph-based ranking method efficiently and effectively approximates the global importances of the subtopics by exploiting the word embedding and content-aware based features. In terms of D#-nDCG@10, *W2V-BGR-JSD* also significantly outperforms the baselines, Kim & Lee (2015), Moreno *et al.* (2014), and Damien *et al.* (2013).

Experiment 3: we evaluated our proposed method *W2V-BGR-JSD* by comparing the performance with the official participants' methods of INTENT-2 and IMINE-2 competitions. The comparative performances of *W2V-BGR-JSD* with the participants of INTENT-2 are demonstrated in Table 4.7. Overall, *W2V-BGR-JSD* outperforms all the official participants' methods in terms of I-rec@10, D-nDCG@10, and D#-nDCG@10 metrics.

With two-tailed paired t-test ($p < 0.05$), in terms of diversity (i.e. I-rec@10), *W2V-BGR-JSD* significantly outperforms all the participants' methods except THUIR-S-E-1A and KLE-S-E-4A, which are statistically indistinguishable from *W2V-BGR-JSD*. In terms of relevance (i.e. D-nDCG@10), *W2V-BGR-JSD* significantly outperforms all the participants' methods except hultech-S-E-1A and THUIR-S-E-4A with two-tailed paired t-test ($p < 0.05$). However, hultech-S-E-1A, THUIR-S-E-1A, and THUIR-S-E-4A are statistically identical with *W2V-BGR-JSD* in terms of D-nDCG@10. In terms of D#-nDCG@10, *W2V-BGR-JSD*

4.5 Experiments and Discussion

Table 4.7: Performance comparison of our proposed method with the official participants of NTCIR-10 INTENT-2 competition at the cutoff rank 10. The best result is in bold. † indicates statistically significant and ◊ indicates statistically indistinguishable from the best.

	Method	I-rec	D-nDCG	D#-nDCG
Our Proposed	<i>W2V-BGR-JSD</i>	0.4745 [†]	0.5394 [†]	0.5048 [†]
NTCIR-10 INTENT-2	THUIR-S-E-4A	0.4364	0.5062 [◊]	0.4713 [◊]
	THCIB-S-E-1A	0.4431	0.4657	0.4544
	THUIR-S-E-1A	0.4512 [◊]	0.4775 [◊]	0.4644 [◊]
	hultech-S-E-1A	0.3680	0.5368 [◊]	0.4524 [◊]
	KLE-S-E-4A	0.4457 [◊]	0.4401	0.4429
	SEM12-S-E-2A	0.3777	0.4290	0.4014
	ORG-S-E-4A	0.3815	0.3829	0.3822
	TUTA1-S-E-1A	0.2181	0.2577	0.2379
	LIA-S-E-4A	0.2000	0.2753	0.2376

significantly outperforms the participants’ methods, however, hultech-S-E-1A, THUIR-S-E-1A, and THUIR-S-E-4A are indistinguishable.

In addition, the comparative performances of *W2V-BGR-JSD* with the official participants of IMINE-2 competitions¹, baseline, and Kim & Lee (2015) are reported in Table 4.8. Overall, *W2V-BGR-JSD* outperforms all the official participants’ methods, baseline, and Kim & Lee (2015) in terms of I-rec@10, D-nDCG@10, and D#-nDCG@10 metrics. With two-tailed paired t-test ($p < 0.05$),

¹<http://www.dl.kuis.kyoto-u.ac.jp/imine2/dataset/#results>

Table 4.8: Performance comparison of our proposed method with the official participants of NTCIR-12 IMINE-2 competition, baseline, and Kim & Lee (2015) at the cutoff rank 10. The best result is in bold. † indicates statistically significant and ◊ indicates statistically indistinguishable from the best.

	Method	I-rec	D-nDCG	D#-nDCG
Our Proposed	<i>W2V-BGR-JSD</i>	0.8235 [†]	0.6911 [†]	0.7581 [†]
NTCIR-12 IMINE-2	KDEIM-Q-E-1S	0.7557	0.6644	0.7101
	HULTECH-Q-E-1Q	0.7280	0.6787 [◊]	0.7033
	RUCIR-Q-E-4Q	0.7601	0.5097	0.6349
Baseline	MergeBGY	0.6144	0.3884	0.5044
Related work	Kim & Lee (2015)	0.5233	0.3210	0.4242

in terms of diversity (i.e. I-rec@10), *W2V-BGR-JSD* shows statistically significant performance over all related methods. In terms of relevance (i.e. D-nDCG@10), *W2V-BGR-JSD* significantly outperforms KDEIM-Q-E-1S, RUCIR-Q-E-4Q, baseline, and Kim & Lee (2015), however, the difference with HULTECH-Q-E-1Q is insignificant. In terms of D#-nDCG@10, *W2V-BGR-JSD* also significantly outperforms the related methods.

Experimental results demonstrate that our proposed bipartite graph based ranking (*BGR*) method with semantic features and *Jensen-Shannon divergence* based novelty function consistently performs better than previous works on both INTENT-2 and IMINE-2 datasets.

4.5.6 Discussion

Taking query "#09 (porteville)" from the INTENT-2 dataset as an example, we listed the top-10 subtopics returned by our proposed *W2V-BGR-JSD*, Baseline (MergeBGY), Kim & Lee (2015), and Moreno *et al.* (2014) in Table 4.9. Note that relevant subtopic is labeled by its intent number. It shows that our proposed *W2V-BGR-JSD* returns 7 relevant subtopics covering 6 intents (out of 7) in the top-10 ranks. Both Baseline (MergeBGY) and Kim & Lee (2015) return 3 relevant subtopics covering 3 intents. Though Moreno *et al.* (2014) returns 5 relevant subtopics, however, those subtopics are redundant and cover only 3 search intents.

The computation bottleneck in our approach is the feature extraction, bipartite graph-based ranking, and diversification. However, using hash tables, features are extracted in linear time. Though equation. (4.11) requires matrix inversion which takes $\mathcal{O}(n^3)$ time, however, computation time is negligible for a few number of candidate subtopics (i.e. small n). Since the diversification problem is NP-hard, the greedy algorithm can achieve the optimal ranking in $\mathcal{O}(n^2)$, which is still negligible for small n (Santos *et al.* (2015)). To satisfy the diverse users, a traditional search engine can be augmented by extending two components: subtopic mining and search diversification. Given a query, a search engine can utilize our method to mine the possible subtopics and

Table 4.9: Results of subtopic ranking for the topic "#09 (porterville)". '-' indicates irrelevant subtopic

Rank	W2V-BGR-JSD	Baseline (MergeBGY)	Kim & Lee (2015)	Moreno <i>et al.</i> (2014)
1	porterville california, 5	porterville recorder, -	porterville college, 1	porterville unified school district, 2
2	porterville college, 1	porterville college, 1	porterville unified school district, 2	porterville community college, 1
3	Map of Porterville CA, 7	porterville unified school district, 2	city of porterville, 5	porterville, -
4	porterville recorder, -	porterville high school, -	porterville recorder, -	porterville developmental center, -
5	city of porterville, 5	city of porterville, 5	porterville, -	porterville homes for sale, -
6	Porterville Jobs, -	porterville police department, -	porterville high school, -	porterville high school, -
7	porterville weather, 4	academy trainings in porterville, -	resident of porterville, -	porterville police department, -
8	Porterville Hotels, 3	resident of porterville, -	porterville city council, -	map of porterville ca, 7
9	porterville mls, -	porterville city council, -	copyright 2005 12 by porterville unified school district, -	porterville adult school, 2
10	porterville adult school, 2	porterville breakfast lions, -	porterville police department, -	porterville college, 1

diversify the initially retrieved top ranked documents based on the mined subtopics.

The limitations of our proposed method are carefully choosing the features, learning the parameters of bipartite graph and diversification, and estimating the novelty of the subtopic. Moreover, optimally combining subtopics from heterogeneous sources might improve the performance.

4.6 Summary

We proposed a method for mining and ranking subtopics of the query. We introduced new features based on word embedding and utilized content-aware features that were selected by a supervised method. The relevance of the candidate subtopic with the query was estimated by introducing a bipartite graph-based ranking (*BGR*) method. For diversifying the candidate subtopics, we introduced a novelty function to estimate the similarity between two candidate subtopics (i.e. short texts) by means of *Jensen-Shannon divergence* through the probability distributions of terms in the retrieved documents from a search engine. We experimented and evaluated our proposed method on NTCIR-10 INTENT-2 and NTCIR-12 IMINE-2 datasets in terms of I-rec, D-nDCG, and D#-nDCG metrics at the cutoff rank 10. We demonstrated that our proposed method significantly outperforms the baselines, the previously known subtopic mining methods (Damien *et al.* (2013); Kim & Lee (2015); Moreno *et al.* (2014)), and the official participants of INTENT-2 and IMINE-2 competitions. In future work, we will incorporate subtopics from different resources in our subtopic mining framework and enhance search result diversification to satisfy the users' information needs. Our other future plan is to organize the subtopics in a multi-level hierarchy and improve the performance of result diversification.

Chapter 5

Bipartite Graph based Ranking of Genetic Disease

With the widespread huge medical knowledge data available on the Internet, it is becoming more and more practical to help doctors in clinical diagnostics by suggesting plausible diseases. Although most of the clinical work is on common diseases, physicians are most likely to search for information when they encounter diagnostic difficulties. Since genetic diseases are difficult to diagnose because of their low prevalence, large number, and broad diversity of symptoms, genetic disease patients are often misdiagnosed or experience long diagnostic delays. In this chapter, we present our proposed bipartite graph based ranking of genetic diseases for a set of clinical phenotypes. In our approach, we have associated a phenotype-gene bipartite graph (\mathcal{PGBG}) with a gene-disease bipartite graph (\mathcal{GDBG}) by producing a phenotype-disease bipartite graph (\mathcal{PDBG}). We have introduced the Bidirectionally-induced Importance Weight (BIW) prediction method to \mathcal{PDBG} for approximating the weight of the edge of disease with phenotype, by considering link information from both sides of the bipartite graph. Experimental results show that our proposed method has outperformed the known related method *Phenomizer* in terms of NDCG@10, NDCG@20, MAP@10, and MAP@20, however, it has performed worse than *Phenomizer* in terms of Kendall's tau-b metric at the top-10 ranks. It also turns out that our proposed method has overall better performance than the baseline methods.

5.1 Introduction

One of the challenging tasks in bioinformatics research is to understand the underlying mechanisms of human disease. There are some genes that are responsible for causing human disease, called disease causative genes (Barrenäs *et al.* (2012)). Phenotypes, the observable characteristics (traits) of an organism, are believed to be determined by genetic materials (DNAs) under environmental influences. In this regard, phenotypes have associations with genes (Carter *et al.* (2013); Yang *et al.* (2011)) and, in turn, causative genes have associations with human diseases (Navlakha & Kingsford (2010); Radivojac *et al.* (2008); Tsafnat *et al.* (2014)) as well. Therefore, there might be a path from a phenotype to human hereditary disease through causative genes with weighting factor along with the edge.

Human diseases might be developed through the phenotypical changes due to some causative genes (Hardy & Singleton (2009); Lechner *et al.* (2012)), and physicians diagnose diseases utilizing their human knowledge of a variety of cases. However, wrong selection of clinical features or medical cases may affect humans severely. Consequently, making the correct diagnosis is unquestionably the most important role of the physician. Many physicians, when faced with difficult cases, rely on general purpose search engines or medical databases (Kortteisto *et al.* (2009); Lombardi *et al.* (2009)). Recent studies have shown that Google Search is the preferred resource for searching medical information (Hider *et al.* (2009); Kortteisto *et al.* (2009); Tang & Ng (2006)), however PubMed, a medical bibliographic search engine is also widely used (Kortteisto *et al.* (2009)). Nevertheless, neither of these systems fits well with the task of finding a diagnosis based on patient data. Google is not optimized for this task, but rather for general web search, whereas PubMed does not rank results by relevance, but merely sorts them by date of publication or other bibliographic information.

Although most of the clinical work is on common diseases, physicians are most likely to search for information when they encounter diagnostic difficulties. Therefore, dealing with such cases is an area where a disease retrieval system could improve the current clinical practice. This is especially important, since

such cases often result in misdiagnosis or diagnosis delay that could negatively affect the patient's outcome (Bouwman *et al.* (2010)). In a complex or even in an unknown case of diseases, physicians may get assistance to make decisions quickly and efficiently. Therefore, a disease retrieval system is an important and supportive tool for physicians.

In our approach, we explore all paths from a phenotype to a disease by utilizing a protein-protein interaction network (*PPIN*), phenotype-gene bipartite graph (*PGBG*), and gene-disease bipartite graph (*GDBG*). In *PPIN*, protein-protein interactions are considered to explore some candidate causative genes, and the explored candidate causative genes are used to extend the gene-disease bipartite graph. In this case, first-neighboring genes of causative genes in *PPIN* are considered to be candidate causative genes. The paths of phenotypes to diseases are examined by associating the phenotype-gene bipartite graph with the extended gene-disease bipartite graph through their common causative genes, and a new phenotype-disease bipartite graph is produced. The weight of an edge of a phenotype with a disease in the phenotype-disease bipartite graph is measured using our proposed genetic disease ranking method *BJW*. Finally, candidate diseases are ranked according to the approximate weight in order to define the most probable diseases for a given set of clinical phenotypes.

The rest of the chapter is organized as follows: **Section 5.2** describes the state of the art. **Section 5.3** includes the general concepts and terminology to comprehend the readers about the contents of this chapter. We introduce the methodology and design of our proposed method in **Section 5.4**. **Section 5.5** includes experiments and evaluation to show the effectiveness of our proposed method. Concluding remarks of our work are described in **Section 5.6**.

5.2 Related Work

Early efforts to use computer diagnostic aids date to more than decade ago (Miller (1994)), but health care institutions have been slow in incorporating them into the clinical workflow. It has been repeatedly asserted in literature that these systems have the potential to reduce diagnostic errors and improve quality of care (Batal *et al.* (2013); Delaney (2008); Kawamoto *et al.* (2005)),

and the utility of some of them was even demonstrated through laboratory evaluation studies (Kawamoto *et al.* (2005)), however few were tested in the field or developed further than the prototype stage, and none of them is in widespread use today.

Recently, there are a number of systems to address the crisis in preventive medicine, where the primary concern is recognizing disease risk and taking action at the earliest signs. Most of these systems are designed to make a prediction about a single disease or a class of some specific diseases. *Phenomizer* is a web-based system that produces a ranked list of hereditary diseases, taking a set of clinical features (Köhler *et al.* (2009)). This system only considers the semantic similarity metrics to measure phenotypic similarity between query phenotypes and disease phenotypes with the use of the Human-Phenotype-Ontology (\mathcal{HPO}) (Robinson & Mundlos (2010)). A long list of possible diseases is presented for a single query by adopting a statistical model to assign p -values to the resulting similarity scores, which is infeasible in real time. However, without considering genetic loci, structural similarity of phenotypes does not always confirm the relevant plausible diseases.

FindZebra (Dragusin *et al.* (2013)) is a vertical search engine specially designed for rare diseases. This system does not consider the genetic effects on disease or phenotypic effects on genes; rather it presents a list of disease documents for a given set of symptoms. *CARE* uses a collaborative filtering method to predict each patient's disease risk based only on their own medical history and that of similar patients (Davis *et al.* (2010)). Moreover, there are some causative genes that may be active in the organism in a different stage of life. Another system named disease interaction prediction (Davis & Chawla (2011)) uses patient medical histories (phenotype data) and known disease-gene association to construct, analyze, and compare disease-disease networks. It provides insight into the interplay between genetics and clinical realities.

In the postgenomic era, it is widely established in bioinformatics and molecular biology to represent associations between biomedical entities as networks, and to analyze their topology to obtain a global understanding of underlying relationships (Barabási *et al.* (2011); Butts (2009); Yıldırım *et al.* (2007)). In this regard, *DisGeNET* is a coherent tool that analyzes and interprets human gene

5.3 General Concepts and Terminologies

network to disease network (Bauer-Mehren *et al.* (2010)). It visualizes the gene-disease association network as a bipartite graph, and provides gene centric and disease centric views of the data.

Interpreting the inherited basis of human disease involves linking genomic variation to clinical phenotypes (Frazer *et al.* (2009)). Establishing this relationship, however, can be challenging for several reasons: the pleiotropy of genes, the genetic heterogeneity of diseases and the limited number of cases (Giallourakis *et al.* (2005)). There are some methods to predict the candidate causative genes. In such, a system (Sun *et al.* (2011)) predicts human disease-related gene cluster using clustering algorithm by integrating protein-protein interaction network and gene expression data, and superimposing a set of known disease genes on human protein-protein interaction network in a different way.

Another system which infers the genotype-phenotype relationship using the Random Walk with Restart algorithm to the Heterogeneous network (*RWRH*) (Li & Patra (2010)), where a heterogeneous network is constructed by connecting the gene network and phenotype network using the phenotype-gene relationship information from the *OMIM* database. However, there is more and more evidence that most human diseases cannot be attributed to single genes but arise due to complex interactions among multiple genetic variants and environmental risk factors (Hirschhorn & Daly (2005)).

5.3 General Concepts and Terminologies

This section introduces the definitions of phenotype, genotype, bipartite graph (bigraph), phenotype-genotype bipartite graph, gene-disease bipartite graph, and *Co-HITS* to comprehend the essence of this chapter.

5.3.1 Phenotype

The phenotype of an organism is the class to which that organism belongs as determined by the description of the physical and behavioral characteristics of the organism (Lewontin (2011)); for example, its morphology, development, biochemical or physiological properties, phenology, behavior, and products

of behavior. The phenotype may change throughout the life of an individual because of environmental changes and the changes associated with aging.

Definition 5.3.1 *Phenotype is the appearance of an individual that results from the interaction of the person's genetic makeup and his or her environment.*

5.3.2 Genotype

The genotype of an organism is the class to which that organism belongs as determined by the description of the actual physical material made up of the DNA that was passed to the organism by its parents at the organism's conception (Lewontin (2011)). The genotype is the descriptor of the genome which is the set of physical DNA molecules inherited from the organism's parents. This is the "internally coded, inheritable information" carried by all living organisms.

Definition 5.3.2 *The genotype is defined as the entire set of genes in a cell, an organism, or an individual.*

5.3.3 Bipartite Graph (Bigraph)

A bipartite graph, also called a bigraph, is a set of graph vertices composed of two disjoint sets such that no two graph vertices within the same set are adjacent. Bipartite graphs are equivalent to two-colorable graphs, and a graph is bipartite if and only if all its cycles are of even length. Consider a bipartite graph $\mathcal{G} = (\mathcal{U} \cup \mathcal{V}, \mathcal{E})$; its vertices can be divided into two disjoint sets \mathcal{U} and \mathcal{V} such that each edge in \mathcal{E} connects a vertex in \mathcal{U} and one in \mathcal{V} ; that is, there is no edge between two vertices in the same set.

Definition 5.3.3 *A bipartite graph is a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ whose vertex set \mathcal{V} can be partitioned into two non-empty sets \mathcal{V}_1 and \mathcal{V}_2 in such a way that every edge in \mathcal{E} of \mathcal{G} joins a vertex in \mathcal{V}_1 to a vertex in \mathcal{V}_2 .*

5.3.4 Phenotype-Genotype Bipartite Graph

If the mechanisms of development were such that every change in genotype resulted in a different phenotype and every different phenotype was the consequence of a difference in genotype, the study of the origin of organic variation would be greatly simplified. Given a knowledge of the phenotype, the underlying causal genotype could be unambiguously inferred, and vice versa. However, the actual correspondence between genotype and phenotype is a many-to-many relation in which any given genotype corresponds to many different phenotypes, and there are different genotypes corresponding to a given phenotype (Lewontin (2011)).

The many-to-many mapping between genotype and phenotype arises from four sources: (1) the relation between the DNA sequence and the chemical structure of proteins; (2) relations between the products of the transcription and translation of the information coded in the genome; (3) the dependence of development and physiology on both the genotype of the organism and the temporal sequence of environments in which the organism develops and functions; and (4) stochastic variations of molecular processes within cells.

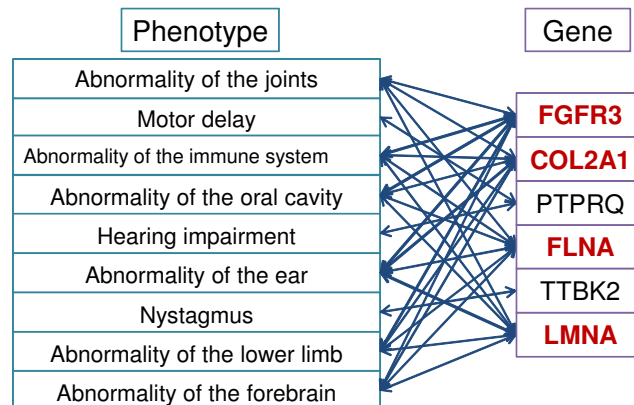


Figure 5.1: The phenotype-gene bipartite graph ($\mathcal{P}\mathcal{G}\mathcal{B}\mathcal{G}$) with unit weight of edge. The red-colored genes are denoted as disease-causative.

For example, in Fig. 5.1, Abnormality of the ear (HP:0000598), a phenotype which is associated with a set of genes such as FGFR3 (2261), COL2A1 (1280), and LMNA (4000). FGFR3 (2261), for example, is a gene which is associated with

a set of phenotypes, including Abnormality of the immune system (HP:0002715), Abnormality of the oral cavity (HP:0000163), and Abnormality of the lower limb (HP:0002814).

Consider a phenotype-gene bipartite graph, $\mathcal{P}\mathcal{G}\mathcal{B}\mathcal{G} := (\mathcal{P} \cup \mathcal{G}, \mathcal{E})$ where $\mathcal{P} = \{p_1, p_2, p_3, \dots, p_m\}$ is the set of m phenotypes, $\mathcal{G} = \{g_1, g_2, g_3, \dots, g_n\}$ is the set of n genes, and every edge in $\mathcal{P}\mathcal{G}\mathcal{B}\mathcal{G}$ joins a phenotype in \mathcal{P} to a gene in \mathcal{G} with unit weight. The $\mathcal{P}\mathcal{G}\mathcal{B}\mathcal{G}$ is depicted in Fig. 5.1 as a bipartite graph where all red-colored genes are denoted as disease-causative.

5.3.5 Gene-Disease Bipartite Graph

The gene-disease associations are represented as a bipartite graph consisting of genes and diseases (Goh *et al.* (2007b); Newman (2003)). Gene and disease nodes are connected through an edge if the according gene-disease association is covered in the gene-disease database. For example, in Fig. 5.2, a set of disease-causative genes FGFR3 (2261), COL2A1 (1280), FLNA (2316), and HTR2A (3356) are associated with a set of human diseases, including Bladder Cancer (OMIM:109800), Diabetes mellitus (OMIM:125853), Colon Cancer (OMIM:114500), Schizophrenia (OMIM:181500), and Heterotopia (OMIM:300049).

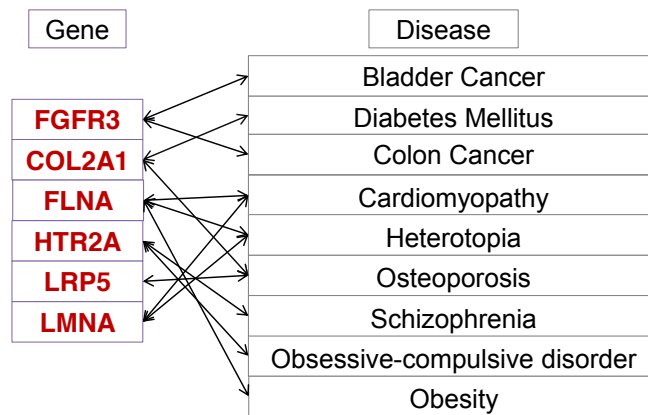


Figure 5.2: The gene-disease bipartite graph ($\mathcal{G}\mathcal{D}\mathcal{B}\mathcal{G}$) with unit weight of edge.

Consider a gene-disease bipartite graph, $\mathcal{G}\mathcal{D}\mathcal{B}\mathcal{G} := (\mathcal{G} \cup \mathcal{D}, \mathcal{E})$ where $\mathcal{G} = \{g_1, g_2, g_3, \dots, g_x\}$ is the set of x causative genes, $\mathcal{D} = \{d_1, d_2, d_3, \dots, d_y\}$ is the set of y diseases, and every edge in $\mathcal{G}\mathcal{D}\mathcal{B}\mathcal{G}$ connects a causative gene in \mathcal{G} to a

disease in \mathcal{D} with unit weight. The \mathcal{GDBG} is depicted in Fig. 5.2 as a bipartite graph.

5.4 Methodology and Design

5.4.1 Data Acquisition

We selected phenotype-gene bipartite graph (\mathcal{PGBG}), Human Phenotype Ontology (\mathcal{HPO}), and disease-phenotypic-annotation from (Robinson & Mundlos (2010))¹. We also selected protein-protein interaction network (\mathcal{PPIN}), and gene-disease bipartite graph (\mathcal{GDBG}) from **Diseasome**² (Goh *et al.* (2007a)). Furthermore, we generated another gene-disease bipartite graph (\mathcal{GDBG}) using *mim2gene* and *morbiditymap* files from **OMIM** (Hamosh *et al.* (2005)). Moreover, there are multiple gene-disease bipartite graphs (\mathcal{GDBG}) across data sources with a difference in gene *ID* or *symbol*. Therefore, a consistent gene-disease bipartite graph was generated by matching all the graphs so that gene *ID* or *symbol* is compatible with the gene *ID* or *symbol* in \mathcal{PGBG} and \mathcal{PPIN} .

In **OMIM**, *omim.txt*³ file has 20,700 disease documents where 5,000 documents contain fields of information, such as clinical synopsis (**CS**), diagnostic process, treatment, number of cases, causative genes, etc. We developed a parser named **OMIM-Parser** to extract phenotype from the **CS** field of each disease document. Given a disease document, our **OMIM-Parser** extracts a phenotype term from the **CS** field, and applies term matching between the extracted phenotype term and the terms available in \mathcal{HPO} for associating with phenotype *ID*, and annotates the matched phenotype with the disease. However, some phenotype terms are contrasted with the terms available in \mathcal{HPO} in the surface form, although they are similar phenotype terms. For example, "Mental retardation, mild to moderate" differs with "Mild to moderate mental retardation", although both are different forms of the same phenotype term. Finally,

¹url: <http://www.human-phenotype-ontology.org/contao/index.php/downloads.html>

²url: <http://diseasome.eu>

³url: <ftp://ftp.ncbi.nih.gov/repository/OMIM/ARCHIVE/omim.txt.Z>

a phenotype-disease annotation file was generated. These phenotype-disease associations were used during the evaluation of the system.

The configurations of the data files used in our system are as follows: in $\mathcal{P}\mathcal{G}\mathcal{B}\mathcal{G}$, there are 6,327 phenotype nodes and 1,807 gene nodes, and $\mathcal{P}\mathcal{P}\mathcal{J}\mathcal{N}$ includes 951 gene nodes. In $\mathcal{G}\mathcal{D}\mathcal{B}\mathcal{G}$, there are 1,271 causative gene nodes and 1,540 disease nodes.

5.4.2 Proposed System Architecture

In this part, the components of our system to estimate the rank of the candidate diseases are described. The architecture of our system is depicted in Fig. 5.3. The system is decomposed into two processing steps: Pre-processing step and Main-Processing step. Pre-processing step includes three sub-processing units: exploring causative genes, associating two bipartite graphs, and generating a weighted data model by estimating the weight of candidate diseases. Main-processing step includes three sub-processing units: collecting the user's query phenotypes, retrieving the relevant diseases from the processed weighted data model, and ranking the retrieved diseases to present a ranked list of possible diseases.

5.4.3 Exploring Causative Genes

In this section, we describe the procedures to explore some candidate causative genes that might be responsible for causing diseases by exploiting $\mathcal{P}\mathcal{P}\mathcal{J}\mathcal{N}$ and $\mathcal{G}\mathcal{D}\mathcal{B}\mathcal{G}$. There are more and more evidences that most human diseases cannot be attributed to a single gene, however they arise due to complex interactions between multiple genetic variants and environmental risk factors (Hirschhorn & Daly (2005)). Since disease-causative genes which are more likely to interact with each other often lead to similar diseases or disorders, a group of genes associated with the similar diseases or disorders should share similar cellular and functional characteristics, as annotated in Gene Ontology ($\mathcal{G}\mathcal{O}$) (Ashburner *et al.* (2000); Goh *et al.* (2007b)). Causative genes which are associated with similar diseases or disorders show an increased tendency for their protein

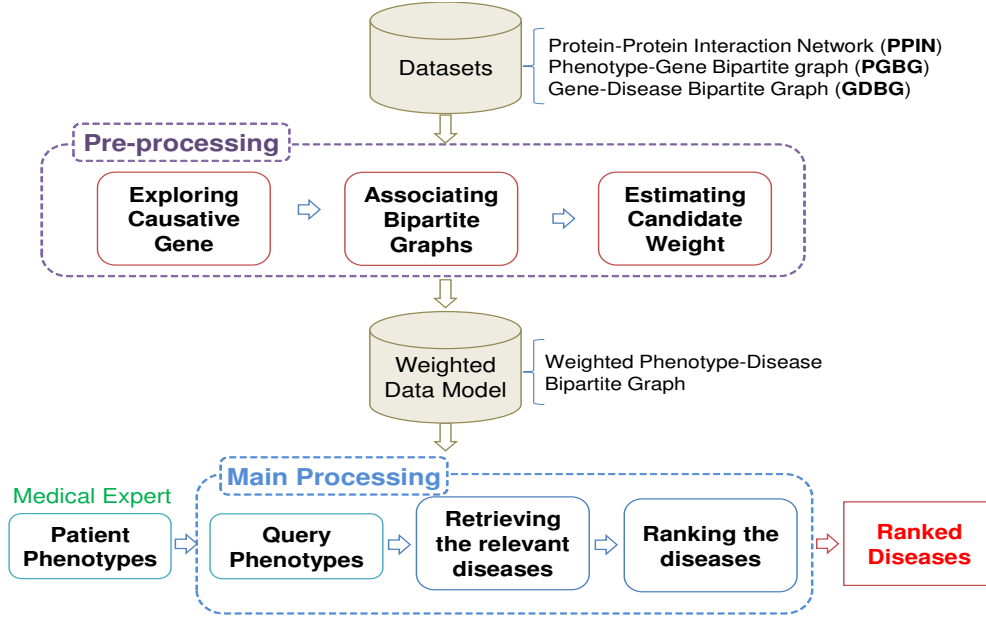


Figure 5.3: Proposed system architecture

products¹ to interact with each other through \mathcal{PPIN} . The working hypothesis of this approach is that the first-neighboring genes of the known causative genes in the \mathcal{PPIN} might be susceptible to diseases i.e. candidate-causative genes. Based on this hypothesis, we explore some candidate-causative genes, and ultimately extend the \mathcal{GDBG} by adding the explored genes and corresponding links with diseases.

The complete procedure of the exploration of candidate causative gene makes use of the \mathcal{PPIN} , and \mathcal{GDBG} graph. It requires three basic operations that are applied to the \mathcal{PPIN} , \mathcal{GDBG} , and extended gene-disease bipartite graph (\mathcal{EGDBG}). The first operation returns a set of causative genes, \mathcal{CG} which are associated with a disease, d in \mathcal{GDBG} . The second operation returns a set of first-neighboring genes, \mathcal{NG} of a causative gene, $cg \in \mathcal{CG}$ from \mathcal{PPIN} . Finally, the third operation updates the \mathcal{EGDBG} by adding the explored candidate gene, $ng \in \mathcal{NG}$ including the corresponding links with the disease, d along with the

¹Note that in the context used here genes and proteins can be used synonymously. Many PPIN do not distinguish isoforms of genes, so that one can say that two genes interact with each other, although actually the gene's products interact with each other.

existing edges of \mathcal{GDBG} .

A simple illustration of a causative gene exploration hypothesis is depicted in Fig. 5.4, where the left network is a sample of the original \mathcal{PPIN} , and the right network is a sample of the \mathcal{PPIN} with explored causative genes. The gene which is the first-neighbor of a causative gene in \mathcal{PPIN} is considered as a candidate-causative gene and denoted as a green color in the explored \mathcal{PPIN} .

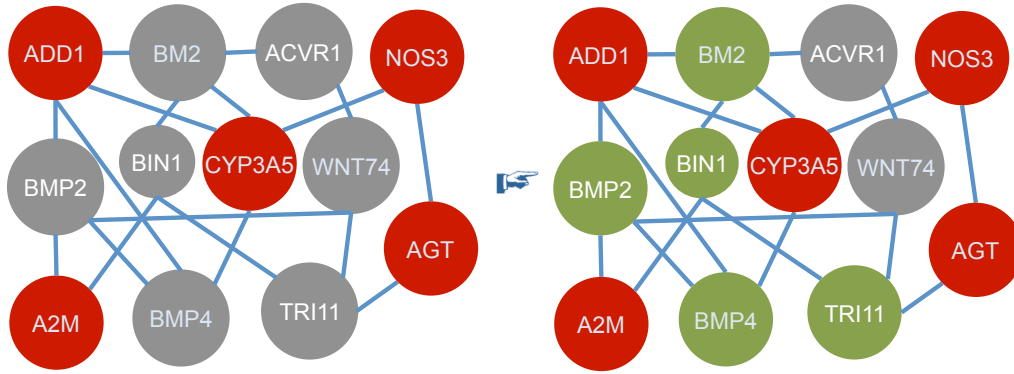


Figure 5.4: A PPIN and an extended PPIN with some explored causative genes in green.

After exploring some candidate-causative genes, it eventually increases more causative gene nodes and edges with diseases in the \mathcal{EGDBG} , which is depicted in Fig. 5.5. For example, "POLG" is a candidate-causative gene denoted by green color, which contributes three new edges ("POLG", "Bladder Cancer"), ("POLG", "Diabetes Mellitus"), and ("POLG", "Osteoporosis") in the \mathcal{EGDBG} . The main goal of this candidate-causative gene exploration is to increase the probability of deciphering more susceptible diseases for a causative gene.

5.4.4 Associating Bipartite Graphs

The method of associating two bipartite graphs is described here. Two bipartite graphs are to be associated if there is a possibility to satisfy the transitive property among the graph nodes. For example, if there are two bipartite graphs $A - B$ and $B - C$, it is possible to associate A with C based on the transitive property among the nodes of A , B , and C , i.e. $a \in A$ is connected to $b \in B$,

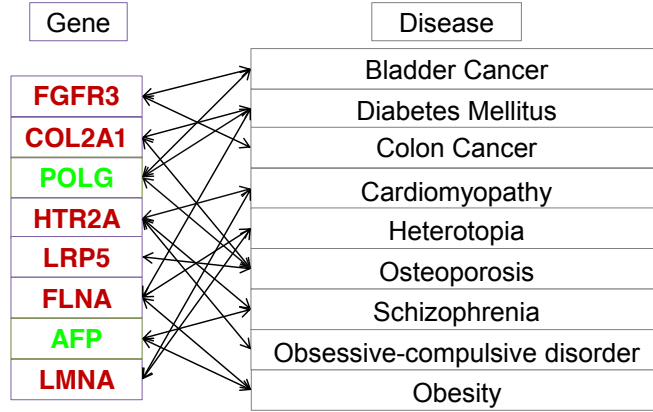


Figure 5.5: An extended gene-disease bipartite graph (\mathcal{EGDBG}) with unit weight of edge. The green are the newly explored candidate-causative genes.

$(a - b)$, and in turn, $b \in B$ is connected to $c \in C$, $(b - c)$; therefore, $a \in A$ can be associated with $c \in C$, $(a - c)$.

It is strongly believed that a phenotype is associated with a set of genes (Mailman *et al.* (2007)), and in turn, a causative gene is also associated with a set of diseases (Khoury *et al.* (2009); Zhang *et al.* (2010)). Therefore, there might be a path from a phenotype to a disease through their associated common causative gene. We count all the paths from a phenotype to a disease through their associated common causative genes, and link the phenotype with the disease in \mathcal{PDBG} , where each edge is labelled with the counted path frequency. Therefore, the frequency label of an edge is the total number of distinct paths from a phenotype to a disease through one or more disease-causative genes.

The procedure for associating two bipartite graphs makes use of the \mathcal{PBG} and \mathcal{EGDBG} . It requires three basic operations. The first operation returns a set of genes that are connected to a phenotype in \mathcal{PBG} . The second operation returns a set of diseases that are connected to a causative gene in \mathcal{EGDBG} . The final operation updates the \mathcal{PDBG} by linking a phenotype to a disease with their counted path frequency.

When associating two bipartite graphs, there are some limitations in our approach. The structure of the associated \mathcal{PDBG} is different from the original \mathcal{PBG} and \mathcal{EGDBG} , which is similar to the one-node-class projection of the classical bipartite graph, i.e. non 1-to-1 correspondence between the original

structure and the associated one. Furthermore, the number of phenotypes in the associated \mathcal{PDBG} is fewer than the number of phenotypes in the original \mathcal{PGBG} and the number of diseases in the associated \mathcal{PDBG} is fewer than the number of diseases in the original \mathcal{EGBG} . This is happening due to the non-smoothness of the \mathcal{PGBG} and \mathcal{EGBG} graphs. For instance, let p be a phenotype node, g be a gene node, and d be a disease node. Let $(a - b)$ denotes an edge. If $(p - g) \in \mathcal{PGBG}$ and $(g - d) \notin \mathcal{EGBG}$, then $(p - d) \notin \mathcal{PDBG}$. Similarly, if $(p - g) \notin \mathcal{PGBG}$ and $(g - d) \in \mathcal{EGBG}$, then $(p - d) \notin \mathcal{PDBG}$. In the above cases, phenotype node p or disease node d might be loosed from the associated \mathcal{PDBG} graph, which is a drawback in our approach. We might reduce these losses of information by introducing some smoothing operation on the original bipartite graph.

As the first-neighboring nodes are targeted for a node in the bipartite graph \mathcal{PGBG} or \mathcal{EGBG} , the information contained by the strength of the co-linked nodes and the edges whose "target" nodes are of degree 1 in the original graph might be loosed from the associated \mathcal{PDBG} graph. To overcome these shortcomings, weight propagation on the bipartite graph \mathcal{PGBG} or \mathcal{EGBG} is essential for boosting the strength of the co-linked nodes and predicting the possible edges before introducing the procedure for associating two bipartite graphs.

5.4.5 Estimating Candidate Weight

The methods of estimating candidate weight of diseases in \mathcal{PDBG} are described here. To estimate the candidate weight of diseases in \mathcal{PDBG} , we propose a ranking method named Bidirectionally-Induced Importance Weight (BIW). We also adopt some well-known weighting method, such as *TF-IDF*, *BM25*, and *JSD* as baselines to compare the performance with our proposed method. Considering the phenotype as term and disease as document, we may have a term-document matrix from the \mathcal{PDBG} . In this regard, we apply *TF-IDF*, *BM25*, and *JSD* methods to the phenotype-disease association matrix. The weighting methods are described in the following sections.

5.4.5.1 Bidirectionally-Induced Importance Weight (BIW) Method

When there are weight propagations on a bipartite graph, we hypothesize that an edge carries information from the nodes of both sides of the bipartite graph. To estimate the weight of an edge in a bipartite graph, we propose a Bidirectionally-induced Importance Weight (BIW) prediction method that uses the link and content information from both sides of the bipartite graph.

Consider a bipartite graph $\mathcal{G} = (\mathcal{U} \cup \mathcal{V}, \mathcal{E})$, where each edge is labelled by a frequency as weight. Given $u_i \in \mathcal{U}$ and $v_j \in \mathcal{V}$, if there is an edge connecting u_i and v_j , the transition probabilities w_{u_i, v_j} and w_{v_j, u_i} are positive, where w_{u_i, v_j} denotes the transition probability from u_i to v_j , and w_{v_j, u_i} denotes the transition probability from v_j to u_i ; otherwise, $w_{u_i, v_j} = w_{v_j, u_i} = 0$. Since the transition probability from state i to all other states must be 1, we have $\sum_{v_j \in \mathcal{V}} w_{u_i, v_j} = 1$ and $\sum_{u_i \in \mathcal{U}} w_{v_j, u_i} = 1$.

For a bipartite graph, there is a natural random walk on the graph with the transition probability as discussed above. Let $\mathcal{W}^{UV} \in \mathcal{R}^{m \times n}$ denote the transition matrix from \mathcal{U} to \mathcal{V} , whose entry (i, j) contains a weight w_{u_i, v_j} from u_i to v_j . Let $\mathcal{W}^{VU} \in \mathcal{R}^{n \times m}$ be the transition matrix from \mathcal{V} to \mathcal{U} , whose entry (j, i) contains a weight w_{v_j, u_i} from v_j to u_i .

Consider a phenotype-disease bipartite graph $\mathcal{PDBG} = (\mathcal{P} \cup \mathcal{D}, \mathcal{E})$, where \mathcal{P} is a set of phenotypes, \mathcal{D} is a set of diseases, and the edges may capture some semantic relations between phenotypes \mathcal{P} and diseases \mathcal{D} . For each edge $(p_i, d_j) \in \mathcal{E}$, we associate a numeric weight f_{ij} , known as the frequency that denotes the number of ways the disease d_j is linked with the phenotype p_i through causative genes.

The transition probability w_{p_i, d_j} from the phenotype p_i to the disease d_j is defined by normalizing the frequency as $w_{p_i, d_j} = \frac{f_{ij}}{\sum_{p_j \in \mathcal{D}} f_{ij}}$, while the transition probability w_{d_j, p_i} from the disease d_j to the phenotype p_i is defined as $w_{d_j, p_i} = \frac{f_{ij}}{\sum_{p_i \in \mathcal{P}} f_{ij}}$. Thus, we can easily obtain the transition matrix \mathcal{W}^{PD} and \mathcal{W}^{DP} .

A sample of the phenotype-disease bipartite graph (\mathcal{PDBG}) is depicted in Fig. 5.6 which illustrates the estimation of weight of a phenotype with a disease. To estimate the weight of an edge (p_i, d_j) of the above \mathcal{PDBG} , we consider the importance weight of both phenotype p_i and disease d_j .

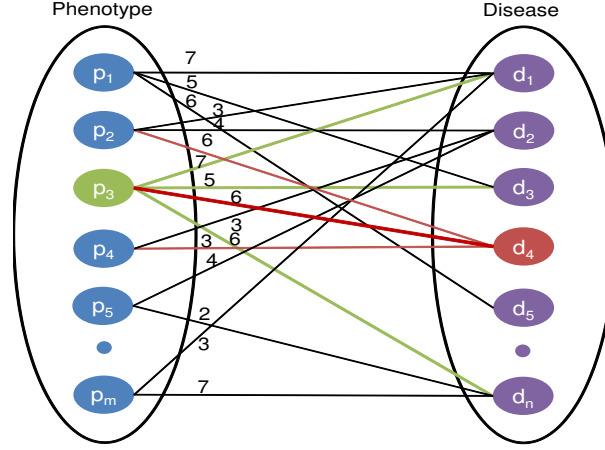


Figure 5.6: A sample of phenotype-disease bipartite graph ($\mathcal{P}\mathcal{D}\mathcal{B}\mathcal{G}$). This figure depicts how the weight between the phenotype p_3 and the disease d_4 is estimated. The phenotype p_3 is connected to the diseases $d_1, d_3, d_4,$ and d_n with frequencies 7, 5, 6, and 6. The disease d_4 is also connected to the phenotypes $p_2, p_3,$ and p_4 with frequencies 7, 6, and 3. The weight of the edge (p_3, d_4) is approximated based on the importance of the phenotype p_3 and the disease d_4 in Equation (5.1).

In order to estimate the importance weight of edge between a phenotype p_i and a disease d_j , we propose a Bidirectionally-induced Importance Weight (BIW) prediction method as follows:

$$weight(p_i, d_j) = \frac{avgl_{\mathcal{D}}}{l_{p_i}} \cdot \frac{w_{d_j, p_i}}{\sum_{p_i \in \mathcal{P}} w_{d_j, p_i}} + \frac{avgl_{\mathcal{P}}}{l_{d_j}} \cdot \frac{w_{p_i, d_j}}{\sum_{d_j \in \mathcal{D}} w_{p_i, d_j}} \quad (5.1)$$

where l_{d_j} is the length of the disease d_j i.e. the number of phenotypes associated with it, $avgl_{\mathcal{D}}$ is the average length of all the diseases in \mathcal{D} , l_{p_i} is the length of the phenotype p_i i.e. the number of diseases associated with it, and $avgl_{\mathcal{P}}$ is the average length of all phenotypes in \mathcal{P} .

In Equation (5.1), $weight(p_i, d_j)$ is the candidate weight of an edge between a phenotype p_i and a disease d_j . The first term of the right side of this equation is the importance weight from a disease d_j to a phenotype p_i , and the second term is the importance weight from a phenotype p_i to a disease d_j . Our BIW method uses the link information from both sides of the bipartite graph to the approximate global candidate weight of every edge of the graph from the local

graph link structure. This method may also distinguish the weight of edges even if they have similar frequencies.

5.4.5.2 TF-IDF Weight

The *TF-IDF* weight of a phenotype p_i on a disease d_j is estimated using the following equation:

$$weight(p_i, d_j) = TF_{p_i, d_j} \cdot IDF_{p_i, \mathcal{D}} \quad (5.2)$$

where *TF* of a phenotype on a disease is defined as follows:

$$TF_{p_i, d_j} = \frac{|p_i \in d_j|}{\sum_k |p_k \in d_j|} \quad (5.3)$$

where p_i is the i^{th} phenotype, d_j is the j^{th} disease, $|p_i \in d_j|$ is the frequency of the edge of phenotype p_i with disease d_j , and $\sum_k |p_k \in d_j|$ is the summation of all the frequencies of edges of phenotypes p_k with the disease d_j .

IDF of a phenotype is defined as follows:

$$IDF_{p_i, \mathcal{D}} = \log \frac{\mathcal{N}}{|\{d \in \mathcal{D} | p_i \in d\}|} \quad (5.4)$$

where \mathcal{N} is the total number of diseases in \mathcal{PDBG} , p_i is the i^{th} phenotype, and $|\{d \in \mathcal{D} | p_i \in d\}|$ is the number of distinct diseases d that are connected to the phenotype p_i in \mathcal{PDBG} .

5.4.5.3 BM25 Weight

The *BM25* weight of a phenotype p_i on a disease d_j is estimated using the following equation:

$$weight(p_i, d_j) = \frac{TF_{p_i, d_j} \cdot (k_1 + 1)}{k_1 \cdot ((1 - b) + (b \cdot \frac{l_d}{avg l_{\mathcal{D}}})) + TF_{p_i, d_j}} \cdot IDF_{p_i, \mathcal{D}} \quad (5.5)$$

where TF_{p_i, d_j} is the *TF* weight of the phenotype p_i on the disease d_j , l_d means the length of disease d i.e. the number of phenotype connected to it, and $avg l_{\mathcal{D}}$

is the average disease length.

IDF of a phenotype p_i in *BM25* Equation (5.5) is defined as follows:

$$IDF_{p_i, \mathcal{D}} = \log \frac{\mathcal{N} - |\{d \in \mathcal{D} | p_i \in d\}| + 0.5}{|\{d \in \mathcal{D} | p_i \in d\}| + 0.5} \quad (5.6)$$

where \mathcal{N} is the total number of diseases, $|\{d \in \mathcal{D} | p_i \in d\}|$ is the summation of all the frequencies of edges of diseases that are connected to the phenotype p_i .

In this model, k_1 and b are free parameters, where we deduce this parameters through empirical evaluation. For this empirical evaluation, we make 81 combinations of values of k_1 and b , whereas $k_1 = \{1.2, 1.3, 1.4, \dots, 2.0\}$ and $b = \{0.50, 0.55, 0.60, \dots, 0.90\}$. For each pair of k_1 and b , we produce the average \mathcal{F} -measure which is depicted in Fig. 5.7. The global peak of the \mathcal{F} -measure curve indicates the optimized value of k_1 and b , which are found to be $k_1 = 1.85$ and $b = 0.82$. This pair of optimized value of k_1 and b is utilized in Equation (5.5).

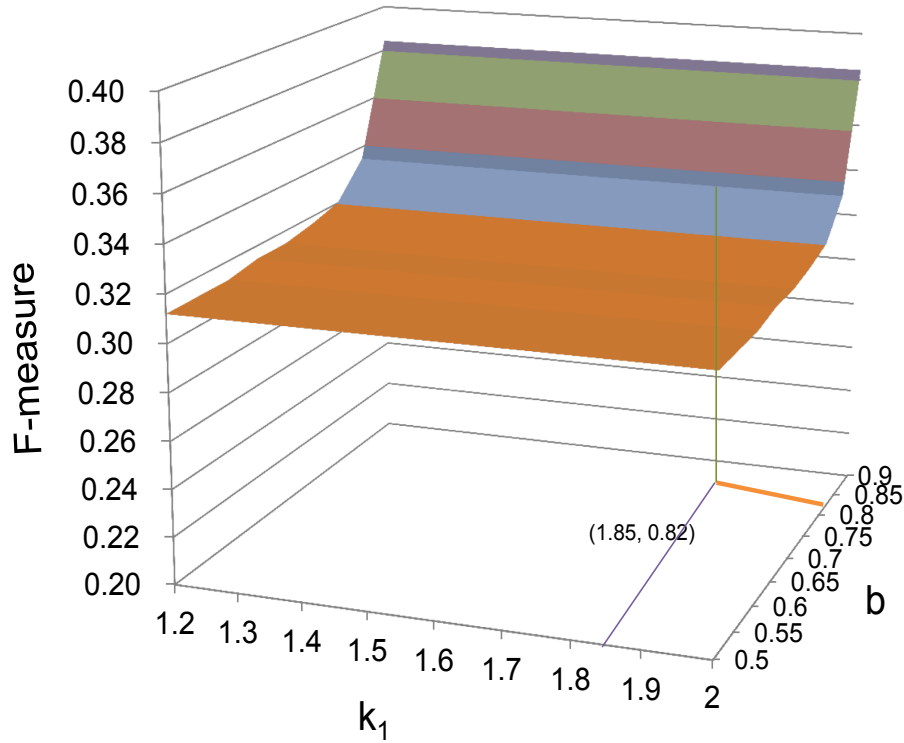


Figure 5.7: Empirical study of *BM25* parameters k_1 and b

5.4.5.4 Jensen-Shannon Divergence (JSD) Weight

The *JSD* weight of a phenotype p_i on a disease d_j is estimated using the following equation:

$$weight(p_i, d_j) = 1.0 - JSD(TF_{p_i, d_j}, IDF_{p_i}) \quad (5.7)$$

where $JSD(TF_{p_i, d_j}, IDF_{p_i})$ is defined as follows:

$$JSD(TF_{p_i, d_j}, IDF_{p_i}) = \frac{1}{2}KL(TF_{p_i, d_j} \parallel M) + \frac{1}{2}KL(M \parallel IDF_{p_i})$$

where $M = \frac{1}{2}(TF_{p_i, d_j} + IDF_{p_i})$, $KL(TF_{p_i, d_j} \parallel M)$ and $KL(M \parallel IDF_{p_i})$ are defined as follows:

$$KL(TF_{p_i, d_j} \parallel M) = TF_{p_i, d_j} \cdot \log\left(\frac{TF_{p_i, d_j}}{M}\right)$$

$$KL(M \parallel IDF_{p_i}) = M \cdot \log\left(\frac{M}{TF_{p_i, d_j}}\right)$$

5.4.5.5 Weighting Phenotype-Disease Bipartite Graph (PDBG)

The candidate weight of an edge is approximated by applying *TF·IDF*, *BM25*, *JSD*, and our proposed method *BJW* individually. The complete procedure for estimating the candidate weight of edges makes use of \mathcal{PDBG} and a specific weight Equation of (5.2), (5.5), (5.7) or (5.1). It requires three basic operations to be applied on the input data. The first operation returns a set of diseases D that are connected to a phenotype p_i in \mathcal{PDBG} . The second operation applies a specific weight Equation of (5.2), (5.5), (5.7) or (5.1) between a phenotype p_i and a disease $d_j \in D$. The third operation updates the weighted phenotype-disease bipartite graph (\mathcal{WPDBG}) by adding the phenotype p_i and disease d_j with the approximate weight. All the above processes are done as a pre-processing step. The main-processing task is outlined in the following section.

5.4.6 Retrieving and Ranking the Diseases

During the main-processing step, given a set of clinical phenotypes by a medical expert (physician), our system retrieves the weighted phenotype-disease bipartite graph (\mathcal{WPDBG}) from the repository as processed data model, and

uses this graph for retrieving the possible diseases that may explain the given set of clinical phenotypes.

Let us assume that a physician chooses a set of clinical phenotypes from the "ABNORMALITY of ABDOMEN" branch of \mathcal{HPO} e.g. splenomegaly (HP:0001744), nausea and vomiting (HP:0002017), atherosclerosis (HP:0002621), abdominal pain (HP:0002027) as patient's query, q . Here, "HP:0001744" is the *id* of a phenotype term splenomegaly in \mathcal{HPO} .

Let us assume that the set of query phenotypes, $q = \{p_1, p_2, \dots, p_k\}$. The specificity of a disease, d to the given query phenotypes, Φ_d is defined as follows:

$$\Phi_d = \sum_{p_i=1}^{|q|} weight(p_i, d) \text{ if } (p_i, d) \in \mathcal{WPD}\mathcal{B}\mathcal{G} \quad (5.8)$$

where $weight(p_i, d)$ is the weight of the edge between the phenotype p_i and the disease d in the $\mathcal{WPD}\mathcal{B}\mathcal{G}$ graph.

By applying Equation (5.8), we may have a list of diseases with their weights. The cumulative weights of disease are estimated, and the diseases are ranked according to their cumulative weights. Through these procedures, we produce a ranked list of hereditary diseases for a given set of clinical phenotypes and present the result to the medical expert.

For example, the top-5 diseases in the ranked list, retrieved by our system based on our proposed model \mathcal{BJW} for the above example query including Alström syndrome (OMIM:203800), Hermansky-pudlak syndrome 1 (OMIM:203300), Sitosteolemia (OMIM:210250), Ovarian hyperstimulation syndrome (OMIM:608115), and Ovarian dysgenesis 1 (OMIM:233300).

Physicians may observe the top-5 or top-10 diseases in the ranked list to diagnose more accurately with the help of our assistive disease estimating system. Physicians may also further monitor clinical phenotypes that are associated with a specific disease for differential diagnosis. The snapshot of our system implementation is depicted in Fig. 5.8. There is an option for choosing the clinical phenotypes and a weight model as input to our system. Given a set

of phenotypes and a weight model, our system may produce a ranked list of probable diseases.

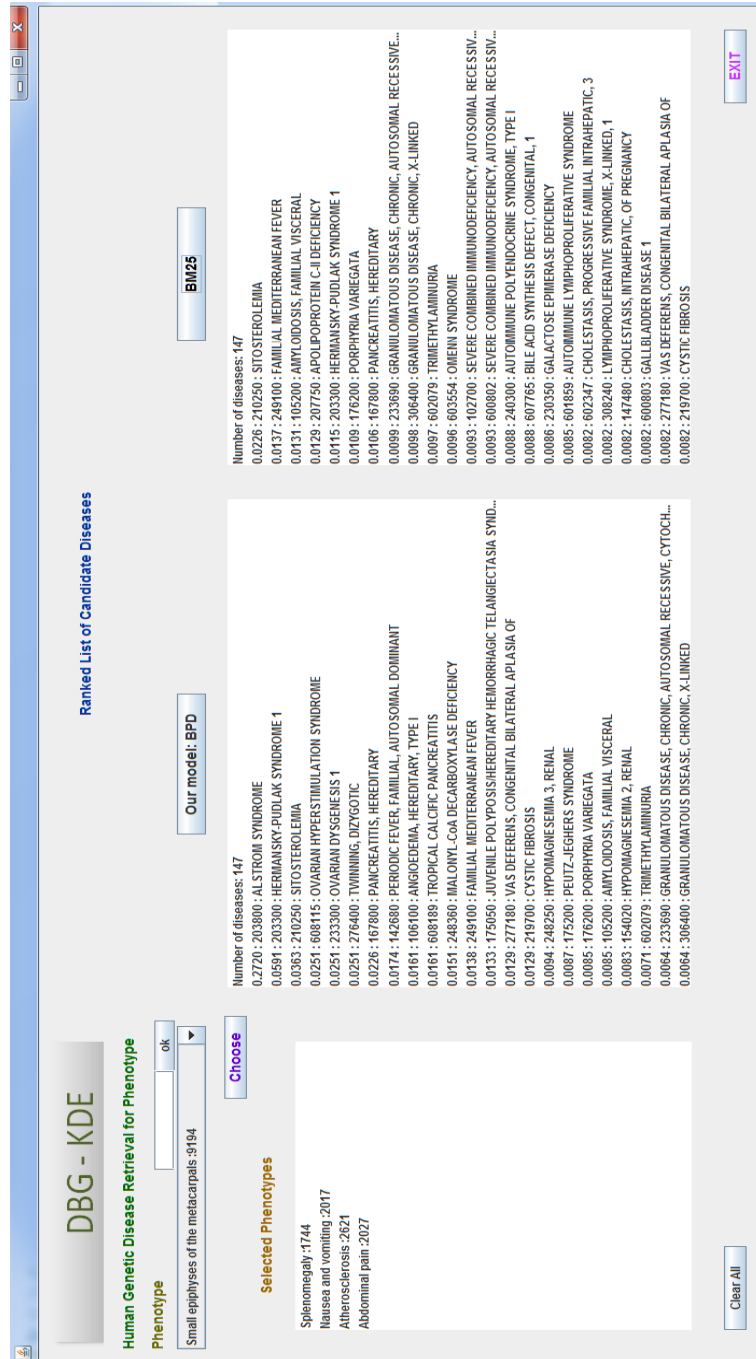


Figure 5.8: An implementation of our disease retrieval system

Table 5.1: A set of queries

No.	Query phenotypes	Possible disease
1	Prominent umbilicus (HP:0001544), Increased hemoglobin (HP:0001900), Esophageal stricture (HP:0002043)	Polycythemia, Erythrocytosis, Gastric sneezing
2	Splenomegaly (HP:0001744), Nausea and vomiting (HP:0002017), Atherosclerosis (HP:0002621), Abdominal pain (HP:0002027)	Lyell syndrome, Enterocolitis, LCAT deficiency
3	Unilateral deafness (HP:0009900), Hyperacusis (HP:0010780), Pulsatile tinnitus (HP:0000361), Abnormal speech discrimination (HP:0001963)	Pyloric atresia, Microcolon, Scleroderma, Kawasaki disease

5.5 Experiments and Evaluation

5.5.1 Query Set

We made a set of 64 queries that were used in the evaluation of the system. Each of these queries are constructed from some specific branch of \mathcal{HPO} , based on the observation that phenotypes in a branch are correlated with each others. In addition, we use two sets of disease-phenotype-annotation as the ground truth collection. A sample of the selected queries with some possible diseases is presented in Table 5.1.

5.5.2 Evaluation Methods

The retrieval accuracies of the proposed method were evaluated using two popular metrics in the field of information retrieval, namely, Mean Average Precision (*MAP*) (Sakai (2005)) and Normalized Discounted Cumulative Gain (*NDCG*) (Järvelin & Kekäläinen (2000)). In addition, pair-size comparisons were carried out to examine the correlation between the retrieval results of the different methods using *Kendall's tau* (Kendall (1938)).

5.5.2.1 Kendall's Tau

The *Kendall's tau* measure is one of the most commonly used measures employed to compute the amount of correlation between two rankings. Given two ranked lists X and Y of length \mathcal{N} , let \mathcal{C} be the total number of concordant pairs (pairs that are ranked in the same order in both rankings) and \mathcal{D} be the total number of discordant pairs (pairs that are ranked in opposite order in the two rankings). The *Kendall's tau* value between the two lists is defined as follows:

$$\tau = \frac{\mathcal{C} - \mathcal{D}}{\mathcal{N}(\mathcal{N} - 1)/2} \quad (5.9)$$

If tied pairs exist on ranks, instead of Equation (5.9), the following measure, called *Kendall's tau-b* (Kendall (1945)), can be used for the computation of associations between ranks:

$$\tau_b = \frac{\mathcal{C} - \mathcal{D}}{\sqrt{(\mathcal{C} + \mathcal{D} + \mathcal{T}_Y)(\mathcal{C} + \mathcal{D} + \mathcal{T}_X)}} \quad (5.10)$$

where \mathcal{T}_X is the number of pairs not tied on rank X , and \mathcal{T}_Y is the number of pairs not tied on rank Y .

5.5.2.2 Evaluation Setup

To evaluate our system in terms of *NDCG*, *MAP* and *Kendall's tau-b*, we run a set of 64 queries in our system and estimate all the evaluation metrics accordingly. Given a query containing a list of phenotypes, our system produces a ranked list of diseases. Each disease in the ranked list is elucidated for its relevance to the given query phenotypes. A disease is considered relevant if it predominantly covers one of the query phenotypes. A disease is also considered relevant if it covers different forms of the query phenotypes i.e phenotype itself or its synonym or even a more general branch of the phenotype. In contrast, a non-relevant disease is one where the query phenotypes cannot be identified after matching all the techniques. That is, the query phenotypes should be connected to the disease directly.

To estimate *NDCG* and *MAP* metrics, we need multilevel relevance grades. We assume 5-level scale relevance grades i.e. irrelevant, marginally relevant, partially relevant, fairly relevant, and highly relevant. Multilevel relevance grades of a disease in the ranked list are measured by using Jaccard’s Index (*JI*) between the set of query phenotypes and the set of phenotypes directly connected with a disease in disease-phenotype-annotation data. Jaccard’s Index (*JI*) is defined as follows:

$$JI = \frac{|DP \cap QP|}{|QP|} \times 100$$

where *DP* (Disease Phenotypes) denotes the set of phenotypes directly connected with a disease *d*, and *QP* (Query Phenotypes) denotes the set of phenotypes appearing in a given query *q*.

The multilevel relevance grade based on Jaccard’s Index (*JI*) is defined as follows:

$$\xi = \begin{cases} 0 & \text{if } JI = 0, \text{ irrelevant} \\ 1 & \text{if } JI \leq 25, \text{ marginally relevant} \\ 2 & \text{if } JI \leq 50, \text{ partially relevant} \\ 3 & \text{if } JI \leq 75, \text{ fairly relevant} \\ 4 & \text{if } JI \leq 100, \text{ highly relevant} \end{cases}$$

5.5.3 Comparison

To compare the effectiveness of our proposed method *BJW* with the statistical methods *TF-IDF*, *BM25*, and *JSD*, we estimate the metrics *NDCG@20* and *MAP@20* for a set of 64 queries, \mathcal{Q} . Given a query $q \in \mathcal{Q}$, a ranked list of hereditary diseases is retrieved by our system based on a specific weighting method i.e. *TF-IDF*, *BM25*, *JSD*, and *BJW* individually. The top-20 diseases in the ranked list are elucidated for their relevancy to the query phenotypes, and multilevel relevance grades are computed. The average *NDCG@20* and *MAP@20* metrics are measured over all queries \mathcal{Q} , which are depicted in Fig. 5.9 (a) and (b), respectively. It turns out in Fig. 5.9 (a) and (b) that our proposed method *BJW* outperforms classical *TF-IDF*, *JSD*, and most of the cases perform better than *BM25*.

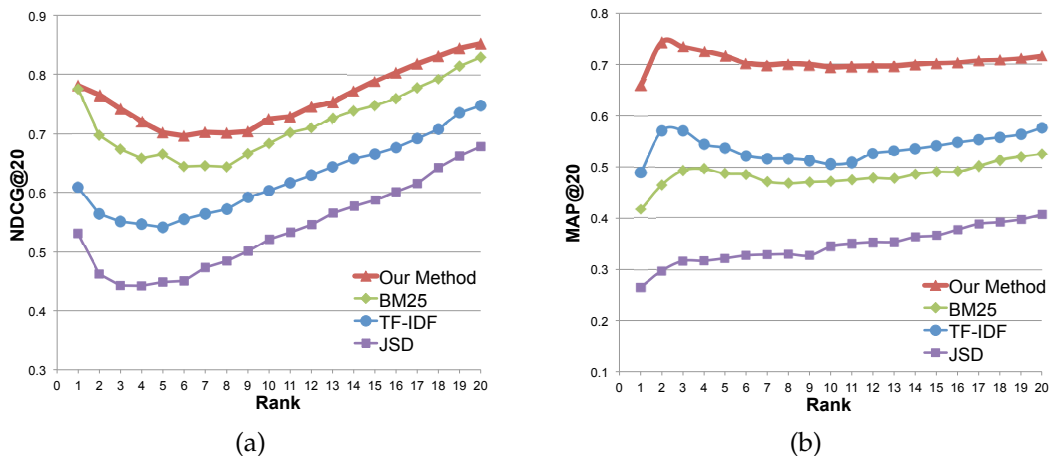


Figure 5.9: The comparison of our proposed method BJW with the baseline methods including $BM25$, $TF-IDF$, and JSD ; (a) $NDCG@20$ metric, (b) $MAP@20$ metric

We evaluate and compare our system's performance with the most known related system named *Phenomizer* (Köhler *et al.* (2009)). In this regard, we estimate $NDCG@10$, $MAP@10$, $NDCG@20$, and $MAP@20$ for a set of 64 queries Q based on our proposed method BJW . However, *Phenomizer* does not produce any type of $NDCG$ or other metrics; it produces a ranked list of genetic diseases given a set of phenotypes. Therefore, we run the same set of 64 queries in the *Phenomizer* system, and estimate the same metrics for the corresponding ranked lists of diseases. The comparison results of our system with *Phenomizer* for $NDCG@10$, $MAP@10$, $NDCG@20$, and $MAP@20$ are depicted in Fig. 5.10 (a), (b), (c), and (d), respectively. It shows in Fig. 5.10 (a) and (b) that our system's performance is to some extent better than *Phenomizer*.

In Fig. 5.10 (c) and (d), it turns out that our system outperforms *Phenomizer* in all aspects. From rank position 1 to 4 and 9 to 20, our system might present more relevant diseases than *Phenomizer*. The top-ranked diseases are very important from the perspective of physicians. Through these comparative experiments, we may ascertain that our system retrieves the most relevant plausible diseases for a given set of phenotypes in *top-10* or *top-20* ranks than *Phenomizer*.

We experiment with and compare our system's performance with link analysis algorithm *Co-HITS* (Deng *et al.* (2009a)). The comparison results of our

5.5 Experiments and Evaluation

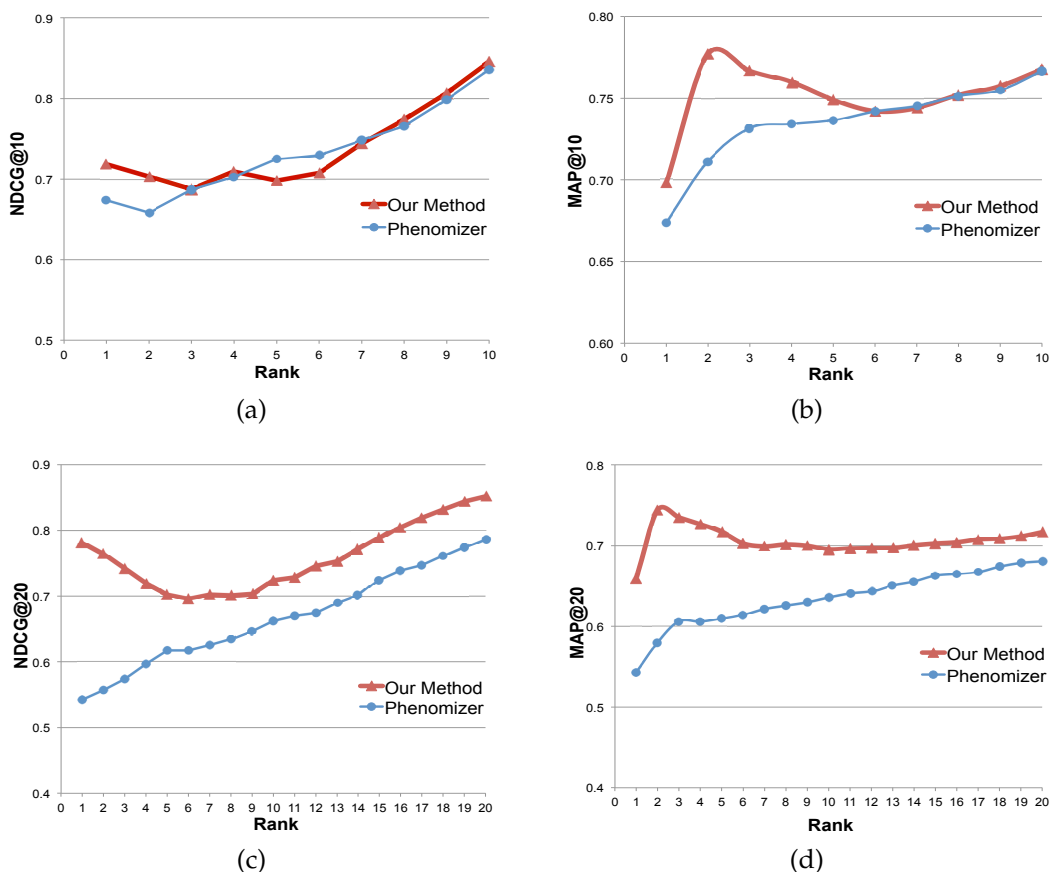


Figure 5.10: The comparison of our proposed method \mathcal{BJW} with $Phenomizer$; (a) $NDCG@10$ metric, (b) $MAP@10$ metric, (c) $NDCG@20$ metric, and (d) $MAP@20$ metric

system with $Co-HITS$ for $NDCG@22$ and $MAP@22$ are depicted in Fig. 5.11 (a) and (b). It turns out that our proposed method \mathcal{BJW} outperforms $Co-HITS$ in disease retrieval problem. In $Co-HITS$, it needs k -times of iterations to estimate the weight of a disease in the graph, however it might lose few importance weights during these iterations.

To observe the correlation between the rankings of two systems, we estimate the *Kendall's tau-b* metric using Equation (5.10) for top-10 diseases. We divide the set of 64 queries into four groups and estimate the average *Kendall's tau-b* metric of \mathcal{BJW} vs. $Ideal$, $Phenomizer$ vs. $Ideal$, and \mathcal{BJW} vs. $Phenomizer$ for each group of queries, and depict the results in Table 5.2. It turns out that $Phenomizer$

5.5 Experiments and Evaluation

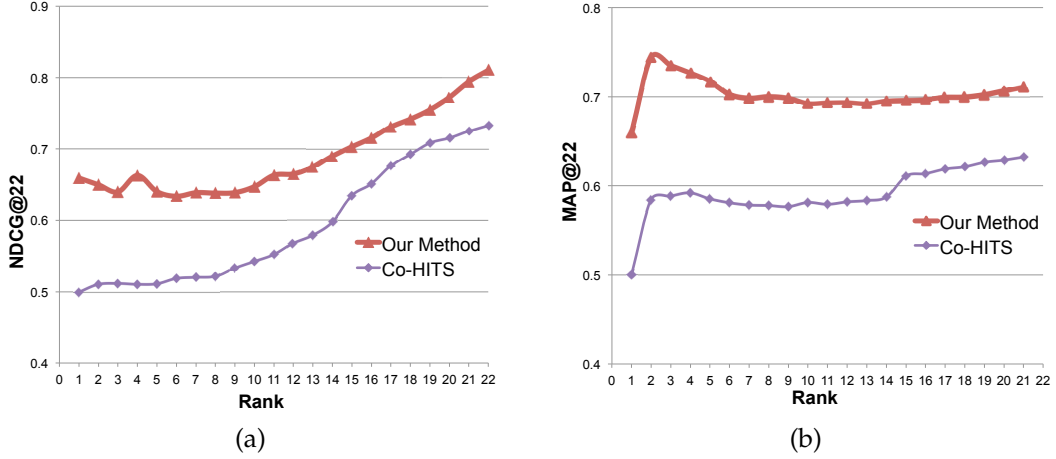


Figure 5.11: The comparison of our proposed method \mathcal{BJW} with link analysis-based algorithm $Co-HITS$; (a) $NDCG@22$ metric, (b) $MAP@22$ metric

Table 5.2: Avg. *Kendall's tau-b* values of our proposed method \mathcal{BJW} with *Phenomizer* for top-10 diseases

Query Group	\mathcal{BJW} vs. Ideal	<i>Phenomizer</i> vs. Ideal	\mathcal{BJW} vs. <i>Phenomizer</i>
(01-16)	0.054	0.102	0.021
(17-34)	0.286	0.401	0.089
(35-48)	0.208	0.266	0.109
(49-64)	0.490	0.111	0.122

outperforms \mathcal{BJW} , by achieving higher *Kendall's tau-b* values as referenced to the ideal ranking of 3 out of 4 query groups. However, the result shows a positive correlation of our proposed method \mathcal{BJW} with *Phenomizer*.

We estimate the average *Kendall's tau-b* metric of our proposed method with the baseline methods, and depict the results in Table 5.3. The ranked list of each weighting method is compared against the ideal rank list individually and then these weighting methods are compared with each other. The result shows that \mathcal{BJW} outperforms other baseline methods as referenced to the ideal ranking of 3 out of 4 query groups, although it performs worse than $BM25$ and JSD for query group (01-16). The positive correlation of \mathcal{BJW} with JSD , and negative correlations with $BM25$ for query group (01-16) and $TF-IDF$ for query group (33-48) have been observed. It also reveals that $BM25$ shows a positive

5.5 Experiments and Evaluation

Table 5.3: Avg. Kendall’s tau-b values of our proposed method \mathcal{BJW} with the baseline methods for top-10 diseases

Query Group	\mathcal{BJW} <i>vs.</i> <i>Ideal</i>	$BM25$ <i>vs.</i> <i>Ideal</i>	$TF-IDF$ <i>vs.</i> <i>Ideal</i>	JSD <i>vs.</i> <i>Ideal</i>	\mathcal{BJW} <i>vs.</i> $BM25$	\mathcal{BJW} <i>vs.</i> $TF-IDF$	\mathcal{BJW} <i>vs.</i> JSD	$BM25$ <i>vs.</i> $TF-IDF$	$BM25$ <i>vs.</i> JSD	$TF-IDF$ <i>vs.</i> JSD
(01-16)	0.054	0.215	0.016	0.246	-0.021	0.096	0.085	0.196	0.183	0.256
(17-32)	0.286	0.127	0.199	0.116	0.044	0.233	0.105	0.240	0.001	-0.026
(33-48)	0.208	0.058	0.058	-0.041	0.057	-0.045	0.125	0.610	-0.038	0.028
(49-64)	0.490	0.444	0.309	0.163	0.220	0.194	0.108	0.313	0.219	0.260

correlation with $TF-IDF$ and a negative correlation with JSD for query group (33-48), whereas $TF-IDF$ shows a negative correlation with JSD for query group (17-32).

Our system retrieves more relevant diseases than *Phenomizer*. We use the genetic overlapping, disease link structure, and phenotype link structure to estimate the weight of a phenotype on a disease. However, *Phenomizer* might only use the structural similarity of clinical phenotypes between the query phenotypes and the phenotypes annotated to a disease using \mathcal{HPO} , which is not always sufficient to estimate the weight of the relevant disease.

There are some other state of the art works e.g. POSSUM (Bankier & Keith (1989)), The London Dysmorphology Database (LDDDB) (Fryns & de Ravel (2002)), as well as the search routine available with the OMIM (Hamosh *et al.* (2005)), Orphanet (Aymé (2003)), and FindZebra (Dragusin *et al.* (2013)). These systems do not provide explicit rankings or measures of plausibility for the potential long lists of candidate diseases. Now, we are trying to compare our system with the rare disease search engine, *FindZebra* (Dragusin *et al.* (2013)).

Although the set of queries \mathcal{Q} were chosen from some specific branches of \mathcal{HPO} , in practical cases, patients only express their relevant clinical features to the physician, who takes some clinical test reports, and might provide the semantically relevant clinical phenotypes to our system. In a practical case, our system will retrieve the most relevant candidate hereditary diseases for a given set of clinical phenotypes.

5.6 Summary

We have proposed a new ranking method for predicting human genetic diseases by associating phenotype-gene with gene-disease bipartite graphs. Our approach is to explore all the paths from a phenotype to a disease through their connected common causative genes, and link the phenotype to the disease with path frequency in a new phenotype-disease bipartite graph (\mathcal{PDBG}). We have introduced the Bidirectionally-induced Importance Weight (\mathcal{BIW}) method to \mathcal{PDBG} for estimating the candidate weights for the edges of phenotypes with diseases, by considering link information from both sides of the bipartite graph. Finally, we have utilized the weighted phenotype-disease bipartite graph (\mathcal{WPDBG}) for retrieving a list of plausible candidate diseases for a given set of clinical phenotypes.

We have experimented with and evaluated our system in terms of $NDCG$, MAP , and Kendall's tau metrics, and demonstrated that our proposed method has outperformed the previously known disease ranking method *Phenomizer* for $NDCG@10$, $MAP@10$, $NDCG@20$, and $MAP@20$, respectively, however, it has performed worse than *Phenomizer* for Kendall's tau-b at the top-10 ranks. We have further conducted comparative experiments of our method with well-known classical statistical methods, including $TF\text{-}IDF$, $BM25$, and JSD . In a few cases, $BM25$ performs better, however, our method outperforms these methods in all aspects. The set of clinical query phenotypes to compute evaluation metrics are chosen from some specific branches of Human-Phenotype-Ontology (\mathcal{HPO}) that might be observed in real cases to patients during diagnosis. In addition, we have conducted comparative experiments of our proposed method \mathcal{BIW} with link analysis-based algorithm *Co-HITS*, and indicated that \mathcal{BIW} performs better than it. Although our evaluations are promising, further validation with the physician is needed to confirm the performance of this method in real diagnosis.

Chapter 6

Conclusions and Future Directions

6.1 Conclusions

In this dissertation, we focus on ranking query subtopics and genetic diseases. Our contributions refer to several significant problems in the field of information retrieval and Bioinformatics. In order to capture the importance of the semantic matching of a query with a subtopic, we have proposed new semantic features based on the locally trained word embedding model. We have also introduced a bipartite graph-based ranking method to estimate the global importance of candidate subtopics by aggregating the local importance of a group of features. To estimate the contextual similarity between a pair of short texts, we have proposed a method of combining a categorical similarity and a mutual information-based similarity by means of *Jensen-Shannon divergence* through the probability distributions of words in the top retrieved documents from a search engine. This contextual similarity is used to estimate the subtopic novelty for result diversification. We have experimented and evaluated our proposed method on NTCIR-10 INTENT-2 and NTCIR-12 IMINE-2 datasets in terms of I-rec@10, D-nDCG@10, and D#-nDCG@10 metrics. We have demonstrated that our proposed method significantly outperforms the baselines, the previously known subtopic mining methods (Damien *et al.* (2013); Kim & Lee (2015); Moreno *et al.* (2014)), and the official participants of INTENT-2 and IMINE-2 competitions.

In the meantime, we have proposed a new ranking method for predicting human genetic diseases for a set of clinical phenotypes. Given two sets of bipartite graphs, we have proposed to associate one bipartite graph (i.e. phenotype-gene) with another bipartite graph (i.e. gene-disease), based on the proximity and the transitive property among the nodes of bipartite graphs. By associating two bipartite graphs, all information is embedded in a new bipartite graph (i.e. phenotype-disease). To estimate the weight of an edge in a bipartite graph, we have introduced the Bidirectionally-induced Importance Weight (BIW) method by considering content and link information from both sides of the bipartite graph. We have experimented with and evaluated our proposed methods in terms of *NDCG*, *MAP*, and Kendall's tau metrics, and demonstrated that our proposed method has outperformed the previously known disease ranking method *Phenomizer* for *NDCG@10*, *MAP@10*, *NDCG@20*, and *MAP@20*, respectively, however, it has performed worse than *Phenomizer* for Kendall's tau-b at the cutoff rank 10. Although our evaluations are promising, further validation with the physician is needed to confirm the performance in real diagnosis.

6.2 Future Directions

6.2.1 Resource based subtopic mining

For subtopic mining, we have exploited the query suggestions provided by the search engines as resources. However, we hypothesize that incorporating subtopics from multiple resources, including query logs, anchor text, and top-K documents in our proposed subtopic mining framework may increase the subtopic recall and boost the accuracy of subtopic mining. These subtopics from diverse sources can enhance the search diversification to satisfy the users' information needs.

6.2.2 Aspect oriented subtopic ranking

There are some subtopics in the final ranked list, which convey semantically similar meaning. That means, subtopics with similar meanings can cover a single aspect or intent of a query. To increase the coverage of the search intent through mining subtopics, aspect-oriented based ranking with diversification might be useful direction. Cluster label of the candidate subtopics can represent an aspect of the query. Embedding the subtopic candidates to estimate the similarity for estimating the novelty, importance, and coverage would be another future direction.

6.2.3 Hierarchical subtopic mining

Given a query, it is possible to generate a multi-level hierarchy of underlying subtopics by analyzing the subtopic mining resources. As for the two-level hierarchy of subtopics, let take the ambiguous query "windows" as an example. The first-level subtopic may be "Microsoft Windows" or "house windows". In the category of "Microsoft Windows", users may be interested in different aspects (second-level subtopics), such as "Windows 10", "Windows update", etc. Therefore, constructing a hierarchical organization organization of subtopics might be useful for understanding the query.

6.2.4 Genetic Disease Ranking

One of our future direction is to apply some link analysis based methods to explore the hidden association of phenotype with diseases. We will experiment with large-scale *PPIN* to explore the candidate-causative gene. Moreover, we will implement semantic similarity metrics between query phenotypes and the disease-annotated phenotypes for re-ranking the ranked list of candidate diseases for differential diagnosis.

6.2.5 Matching Diseases and Phenotypes Ontologies

In the association of diseases with phenotypes, there are some wrong aligned pairs, which might decrease the performance of the retrieval of diseases. To produce more accurate alignment pairs of disease with phenotypes, one can introduce ontology alignment algorithm on the disease ontology, gene ontology, and phenotype ontology. In this direction, we will experiment with the dataset in the evaluation forum of the OAEI¹ workshop in the disease-phenotype ontology alignment task.

¹<http://oaei.ontologymatching.org/2016>

7 Related Publications

Articles in International Journals:

- Ullah, M. Z., Aono, M. : A Bipartite Graph-based Ranking Approach to Query Subtopics Diversification Focused on Word Embedding Features, IEICE Transactions on Information and Systems (IEICE TOIS), 11 pages, pp. 3090-3100, Volume E99-D, Issue 12, 2016, DOI: 10.1587/transinf.2016-EDP7190.
- Ullah, M. Z., Aono, M., Seddiqui, M. H. : Estimating a Ranked List of Human Genetic Diseases by Associating Phenotype-Gene with Gene-Disease Bipartite Graphs, ACM Transactions on Intelligent System and Technology (ACM TIST), 21 pages, pp. 56:1 56:21, Volume 6, Issue 4, 2015, DOI: 10.1145/2700487.

Articles in International Conferences:

- Ullah, M. Z., Shajalal, M., Chy, A. N., Aono, M. : Query Subtopic Mining Exploiting Word Embedding for Search Result Diversification, Twelfth Asia Information Retrieval Societies Conference (AIRS 2016), pp. 308-314, November 30-December 2, Beijing, China, 2016, DOI:10.1007/978-3-319-48051-0_24. (**Best Presentation Award**)
- Shajalal, M., Ullah, M. Z., Chy, A. N., Aono, M. : Query Subtopic Diversification based on Cluster Ranking and Semantic Features, IEEE Inter-

national Conference on Advanced Informatics: Concepts, Theory and Application (ICAICTA 2016), 6 pages, August 13-16, Penang, Malaysia, 2016. (**In Press**)

- Chy, A. N., Ullah, M. Z., Aono, M. : Combining Temporal and Content Aware Features for Microblog Retrieval, IEEE International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA 2015), pp. 1-6, August 19-22, Chonburi, Thailand, 2015, DOI:10.1109/ICAICTA.2015.7335353. (**Best Student Paper Award**)
- Ullah, M. Z., Aono, M. : Query Subtopic Mining for Search Result Diversification, IEEE International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA 2014), pp. 309-314, August 20-21, Bandung, Indonesia, 2014, DOI:10.1109/ICAICTA.2014.7005960.
- Reshma, I. A., Ullah, M. Z., Aono, M., Ontology based Classification for Multi-label Image Annotation, IEEE International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA 2016), pp. 226-231, August 20-21, Bandung, Indonesia, 2014, DOI:10.1109/ICAICTA.2014.7005945. (**Best Student Paper Award**)
- Ullah, M. Z., Aono, M., Seddiqui, M. H. ; Estimating a Ranked List of Human Hereditary Diseases for Clinical Phenotypes by Using Weighted Bipartite Networks, 35th IEEE International Conference on Engineering in Medicine and Biology Society (EMBS), pp. 3475-3478, July 7, Osaka, Japan, 2013, DOI:10.1109/EMBC.2013.6610290

Workshop/Technical Papers

- Ullah, M. Z., Aono, M. : KDEIR at CLEF eHealth 2016: Health Documents Re-ranking Based on Query Variations, CLEF 2016 Evaluation Labs and Workshop, Online Working Notes (CEUR), pp. 167-170, vol. 1609, September 5-8, Évora, Portugal, 2016.

-
- **Ullah, M. Z.**, Shajalal, M., Aono, M. : KDEIM at NTCIR-12 IMine-2 Search Intent Mining Task: Query Understanding through Diversified Ranking of Subtopics, Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies, NTCIR, pp. 60–63, June 7-10, Tokyo, Japan, 2016.
 - Chy, A. N., **Ullah, M. Z.**, Shajalal, M., Aono, M. : KDETm at NTCIR-12 Temporalia Task: Combining a Rule-based Classifier with Weakly Supervised Learning for Temporal Intent Disambiguation, Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies, NTCIR, pp. 281-284, June 7-10, Tokyo, Japan, 2016.
 - **Ullah, M. Z.**, Aono, M. : KDEVIR at ImageCLEF 2015 Scalable Image Annotation, Localization, and Sentence Generation task: Ontology based Multilabel Image Annotation, Proceedings of the CLEF, 6 pages, September 8-11, Toulouse, France, 2015.
 - Chy, A. N., **Ullah, M. Z.**, Aono, M. : A Time and Context Aware Re-ranker for Microblog Retrieval, The 29th Annual Conference of the Japanese Society for Artificial Intelligence, May 30 - June 2, Hokkaido, Japan, 2015.
 - **Ullah, M. Z.**, Aono, M. : SEM13 at the NTCIR-11 IMINE Task: Subtopic Mining and Document Ranking Subtasks, In Proceedings of the NTCIR-11, pp. 64-68, December 9-12, Tokyo, Japan, 2014.
 - Reshma, I. A., **Ullah, M. Z.**, Aono, M. : KDEVIR at ImageCLEF 2014 Scalable Concept Image Annotation Task: Ontology based Automatic Image Annotation, Conference and Labs of the Evaluation Forum, CLEF 2014, 9 pages, September 15-18, Sheffield, UK, 2014. (**Winning Team**)
 - Reshma, I. A., **Ullah, M. Z.**, Aono, M. : Ontology based Supervised Learning for Image Annotation, The Institute of Electronics, Information and Communication Engineers (IEICE), Image Engineering (IE), IE2014-22-IE2014-29, Technical Report, Vol. 114, No.172, PP. 41-56, August 1, Chiba, Japan, 2014.

-
- **Ullah, M. Z.**, Aono, M. : Query Suggestions with Global Consistency on User Click Graph, In Proceedings of the WI2 Web Intelligence and Interaction, ARG SIG-WI2, WI2-2013-23, PP. 15-20, December 13-14, Yokohama, Japan, 2013.
 - Reshma, I. A., **Ullah, M. Z.**, Aono, M. : KDEVIR at ImageCLEF 2013 Image Annotation Task, Conference and Labs of the Evaluation Forum, CLEF 2013, September 23-26, Valencia, Spain, 2013.
 - **Ullah, M. Z.**, Aono, M., Seddiqui, H.: SEM12 at the NTCIR-10 INTENT-2 English Subtopic Mining Subtask, In Proceeding of NTCIR-10, NTCIR, June 18-21, Tokyo, Japan, 2013.

References

- AGRAWAL, R., GOLLAPUDI, S., HALVERSON, A. & IEONG, S. (2009). Diversifying search results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, 5–14, ACM. [3, 9](#)
- AMATI, G. (2003). *Probability models for information retrieval based on divergence from randomness*. Ph.D. thesis, University of Glasgow. [36](#)
- AMATI, G., AMODEO, G., BIANCHI, M., GAIBISSO, C. & GAMBOSI, G. (2008). Fub, iasi-cnr and university of tor vergata at trec 2008 blog track. Tech. rep., DTIC Document. [18](#)
- ASHBURNER, M., BALL, C., BLAKE, J., BOTSTEIN, D., BUTLER, H., CHERRY, J., DAVIS, A., DOLINSKI, K., DWIGHT, S., EPPIG, J. *et al.* (2000). Gene ontology: tool for the unification of biology. *Nature Genetics*, **25**, 25. [66](#)
- AYMÉ, S. (2003). Orphanet, an information site on rare diseases. *Soins; la revue de référence infirmière*, **672**, 46–47. [84](#)
- BANKIER, A. & KEITH, C. (1989). Short communication: Possum: The micro-computer laser-videodisk syndrome information system. *Ophthalmic Genetics*, **10**, 51–52. [84](#)
- BARABÁSI, A., GULBAHCE, N. & LOSCALZO, J. (2011). Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, **12**, 56–68. [11, 60](#)

REFERENCES

- BARABÁSI, A.L., JEONG, H., NÉDA, Z., RAVASZ, E., SCHUBERT, A. & VICSEK, T. (2002). Evolution of the social network of scientific collaborations. *Physica A: Statistical mechanics and its applications*, **311**, 590–614. [10](#)
- BARRENÁS, F., CHAVALI, S., ALVES, A., COIN, L., JARVELIN, M., JÖRNSTEN, R., LANGSTON, M., RAMASAMY, A., ROGERS, G., WANG, H. *et al.* (2012). Highly interconnected genes in disease-specific networks are enriched for disease-associated polymorphisms. *Genome Biology*, **13**, R46. [58](#)
- BATAL, I., VALIZADEGAN, H., COOPER, G.F. & HAUSKRECHT, M. (2013). A temporal pattern mining approach for classifying electronic health record data. *ACM Trans. Intell. Syst. Technol.*, **4**, 63:1–63:22. [60](#)
- BAUER-MEHREN, A., RAUTSCHKA, M., SANZ, F. & FURLONG, L. (2010). DisGeNET: a Cytoscape plugin to visualize, integrate, search and analyze gene-disease networks. *Bioinformatics (Oxford, England)*, **26**, 2924–2926. [11](#), [61](#)
- BENDERSKY, M., CROFT, W.B. & DIAO, Y. (2011). Quality-biased ranking of web documents. In *Proceedings of the fourth ACM international conference on Web search and data mining*, 95–104, ACM. [24](#), [36](#)
- BENJAMINS, R., CONTRERAS, J., CORCHO, O. & GOMEZ-PEREZ, A. (2002). The six challenges of the semantic web. In *Proceedings of the Eighth International Conference on Principles of Knowledge Representation and Reasoning, KR2002*, Morgan Kaufmann Publishers. [25](#)
- BERNERS-LEE, T., FISCHETTI, M. & FOREWORD BY-DERTOUZOS, M.L. (2000). *Weaving the Web: The original design and ultimate destiny of the World Wide Web by its inventor*. HarperInformation. [24](#)
- BOGERS, T. (2010). Movie recommendation using random walks over the contextual graph. In *Proc. of the 2nd Intl. Workshop on Context-Aware Recommender Systems*. [10](#)
- BOLDI, P., BONCHI, F., CASTILLO, C., DONATO, D., GIONIS, A. & VIGNA, S. (2008). The query-flow graph: model and applications. In *Proceedings of the*

REFERENCES

- 17th ACM conference on Information and knowledge management*, 609–618, ACM. [7, 29](#)
- BOLLACKER, K., EVANS, C., PARITOSH, P., STURGE, T. & TAYLOR, J. (2008). Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, 1247–1250, ACM. [3](#)
- BOUWMAN, M.G., TEUNISSEN, Q.G., WIJBURG, F.A. & LINTHORST, G.E. (2010). Doctor google ending the diagnostic odyssey in lysosomal storage disorders: parents using internet search engines as an efficient diagnostic strategy in rare diseases. *Archives of Disease in Childhood*, **95**, 642–644. [59](#)
- BRODER, A. (2002). A taxonomy of web search. *ACM Sigir forum*, **36**, 3–10. [7, 29](#)
- BUTTS, C. (2009). Revisiting the foundations of network analysis. *Science*, **325**, 414. [11, 60](#)
- CALLAN, J., HOY, M., YOO, C. & ZHAO, L. (2009). Clueweb09 data set. [44](#)
- CAO, L., GUO, J. & CHENG, X. (2011). Bipartite graph based entity ranking for related entity finding. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2011 IEEE/WIC/ACM International Conference on*, 130–137, IEEE. [10, 38](#)
- CARBONELL, J. & GOLDSTEIN, J. (1998). The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, 335–336, ACM. [9, 28](#)
- CARTER, H., HOFREE, M. & IDEKER, T. (2013). Genotype to phenotype via network analysis. *Current Opinion in Genetics & Development*, **23**, 611–621. [58](#)
- CARTERETTE, B. (2011). An analysis of np-completeness in novelty and diversity ranking. *Information Retrieval*, **14**, 89–106. [3, 9](#)

REFERENCES

- CLARKE, C.L., KOLLA, M., CORMACK, G.V., VECHTOMOVA, O., ASHKAN, A., BÜTTCHER, S. & MACKINNON, I. (2008). Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, 659–666, ACM. [9](#), [27](#)
- CLARKE, C.L., CRASWELL, N. & SOBOROFF, I. (2009). Overview of the trec 2009 web track. Tech. rep., DTIC Document. [7](#), [29](#)
- DAMIEN, A., ZHANG, M., LIU, Y. & MA, S. (2013). Improve web search diversification with intent subtopic mining. In *Natural Language Processing and Chinese Computing*, 322–333, Springer. [8](#), [30](#), [45](#), [50](#), [51](#), [55](#), [86](#)
- DANG, V. & CROFT, W.B. (2012). Diversity by proportionality: an election-based approach to search result diversification. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, 65–74, ACM. [9](#)
- DAVIS, D. & CHAWLA, N. (2011). Exploring and exploiting disease interactions from multi-relational gene and phenotype networks. *PloS One*, **6**, e22670. [60](#)
- DAVIS, D., CHAWLA, N., CHRISTAKIS, N. & BARABÁSI, A. (2010). Time to CARE: a collaborative engine for practical disease prediction. *Data Mining and Knowledge Discovery*, **20**, 388–415. [60](#)
- DELANEY, B.C. (2008). Potential for improving patient safety by computerized decision support systems. *Family practice*, **25**, 137–138. [60](#)
- DENG, H., LYU, M.R. & KING, I. (2009a). A generalized algorithm and its application to bipartite graphs. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 239–248, ACM, ACM, New York, NY, USA. [10](#), [22](#), [82](#)
- DENG, H., LYU, M.R. & KING, I. (2009b). A generalized co-hits algorithm and its application to bipartite graphs. In *Proceedings of the 15th ACM SIGKDD*

REFERENCES

- International Conference on Knowledge Discovery and Data Mining, KDD '09*, 239–248, ACM, New York, NY, USA. [39](#)
- DRAGUSIN, R., PETCU, P., LIOMA, C., LARSEN, B., JØRGENSEN, H.L., COX, I.J., HANSEN, L.K., INGWERSEN, P. & WINTHER, O. (2013). Findzebra: A search engine for rare diseases. *International Journal of Medical Informatics*, **82**, 528– 538. [60](#), [84](#)
- FELLBAUM, C. (1998). *WordNet*. Wiley Online Library. [21](#)
- FRAZER, K., MURRAY, S., SCHORK, N. & TOPOL, E. (2009). Human genetic variation and its contribution to complex traits. *Nature Reviews Genetics*, **10**, 241–251. [61](#)
- FRYNS, J. & DE RAVEL, T. (2002). London dysmorphology database, london neurogenetics database and dysmorphology photo library on cd-rom [version 3] 2001. *Human Genetics*, **111**, 113–113. [84](#)
- GANGULY, D., ROY, D., MITRA, M. & JONES, G.J. (2015). Word embedding based generalized language model for information retrieval. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 795–798, ACM. [11](#)
- GIALLOURAKIS, C., HENSON, C., REICH, M., XIE, X. & MOOTHA, V. (2005). Disease gene discovery through integrative genomics. *Annual Review of Genomics and Human Genetics*, **6**, 381. [61](#)
- GOH, K., CUSICK, M., VALLE, D., CHILDS, B., VIDAL, M. & BARABÁSI, A. (2007a). “the human disease network (the human disesome)”. [65](#)
- GOH, K.I., CUSICK, M.E., VALLE, D., CHILDS, B., VIDAL, M. & BARABÁSI, A.L. (2007b). The human disease network. *Proceedings of the National Academy of Sciences*, **104**, 8685–8690. [64](#), [66](#)
- GRUBER, T.R. (1995). Toward principles for the design of ontologies used for knowledge sharing? *International journal of human-computer studies*, **43**, 907–928. [24](#)

REFERENCES

- HAMOSH, A., SCOTT, A., AMBERGER, J., BOCCHINI, C. & MCKUSICK, V. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, **33**, D514. [65](#), [84](#)
- HARDY, J. & SINGLETON, A. (2009). Genomewide association studies and human disease. *New England Journal of Medicine*, **360**, 1759–1768. [58](#)
- HIDER, P.N., GRIFFIN, G., WALKER, M. & COUGHLAN, E. (2009). The information seeking behavior of clinical staff in a large health care organization. *Journal of the Medical Library Association: JMLA*, **97**, 47. [58](#)
- HIRSCHHORN, J. & DALY, M. (2005). Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics*, **6**, 95–108. [61](#), [66](#)
- HU, J., WANG, G., LOCHOVSKY, F., SUN, J.T. & CHEN, Z. (2009). Understanding user's query intent with wikipedia. In *Proceedings of the 18th international conference on World wide web*, 471–480, ACM. [8](#), [29](#)
- HU, S., DOU, Z., WANG, X., SAKAI, T. & WEN, J.R. (2015). Search result diversification based on hierarchical intents. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, 63–72, ACM. [3](#), [8](#), [27](#), [30](#)
- JÄRVELIN, K. & KEKÄLÄINEN, J. (2000). Ir evaluation methods for retrieving highly relevant documents. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 41–48, ACM, ACM, New York, NY, USA. [78](#)
- KAWAMOTO, K., HOULIHAN, C.A., BALAS, E.A. & LOBACH, D.F. (2005). Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. *BMJ*, **330**, 765. [60](#)
- KENDALL, M.G. (1938). A new measure of rank correlation. *Biometrika*, **30**, 81–93. [78](#)

REFERENCES

- KENDALL, M.G. (1945). The treatment of ties in ranking problems. *Biometrika*, **33**, 239–251. [79](#)
- KENTER, T. & DE RIJKE, M. (2015). Short text similarity with word embeddings. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, 1411–1420, ACM. [11](#)
- KHOURY, M., BEDROSIAN, S., GWINN, M., HIGGINS, J., IOANNIDIS, J. & LITTLE, J. (2009). *Human Genome Epidemiology, 2nd Edition: Building the evidence for using genetic information to improve health and prevent disease*. Oxford University Press, USA. [69](#)
- KIM, S.J. & LEE, J.H. (2013). The kle’s subtopic mining system for the ntcir-10 intent-2 task. In *Proceedings of the 10th NTCIR Conference on Evaluation of Information Access Technologies*, NTCIR. [9](#), [31](#)
- KIM, S.J. & LEE, J.H. (2015). Subtopic mining using simple patterns and hierarchical structure of subtopic candidates from web documents. *Inf. Process. Manage.*, **51**, 773–785. [8](#), [30](#), [45](#), [50](#), [51](#), [52](#), [53](#), [54](#), [55](#), [86](#)
- KÖHLER, S., SCHULZ, M., KRAWITZ, P., BAUER, S., DÖLKEN, S., OTT, C., MUNDLOS, C., HORN, D., MUNDLOS, S. & ROBINSON, P. (2009). Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *The American Journal of Human Genetics*, **85**, 457–464. [60](#), [81](#)
- KORTTEISTO, T., KAILA, M., KUNNAMO, I., NYBERG, P., AALTO, A.M. & RISSANEN, P. (2009). Self-reported use and clinical usefulness of second-generation decision support—a survey at the pilot sites for evidence-based medicine elec-tronic decision support (ebmeds). *Finnish Journal of eHealth and eWelfare*, **1**, 161–169. [58](#)
- KROVETZ, R. (1993). Viewing morphology as an inference process. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, 191–202, ACM. [21](#), [33](#)

REFERENCES

- LAFFERTY, J. & ZHAI, C. (2001). Document language models, query models, and risk minimization for information retrieval. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, 111–119, ACM. [36](#)
- LECHNER, M., HOHN, V., BRAUNER, B., DUNGER, I., FOBO, G., FRISHMAN, G., MONTRONE, C., KASTENMULLER, G., WAEGELE, B. & RUEPP, A. (2012). Cider: multifactorial interaction networks in human diseases. *Genome Biology*, **13**, R62. [58](#)
- LEWONTIN, R. (2011). The genotype/phenotype distinction. In E.N. Zalta, ed., *The Stanford Encyclopedia of Philosophy*, Stanford CA: Metaphysics Research Lab, Stanford, USA, summer 2011 edn. [62](#), [63](#)
- LI, Y. & PATRA, J. (2010). Genome-wide Inferring Gene–Phenotype Relationship by Walking on the Heterogeneous Network. *Bioinformatics*, **26**, 1219. [11](#), [61](#)
- LIANG, S., REN, Z. & DE RIJKE, M. (2014). Fusion helps diversification. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, 303–312, ACM. [10](#)
- LIN, J. (1991). Divergence measures based on the shannon entropy. *Information Theory, IEEE Transactions on*, **37**, 145–151. [20](#)
- LIU, T.Y. (2009). Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, **3**, 225–331. [34](#)
- LIU, Y., SONG, R., ZHANG, M., DOU, Z., YAMAMOTO, T., KATO, M.P., OHSHIMA, H. & ZHOU, K. (2014). Overview of the ntcir-11 imine task. In *Proceedings of the 11th NTCIR Conference on Evaluation of Information Access Technologies*, NTCIR. [3](#), [27](#)
- LOMBARDI, C., GRIFFITHS, E., MCLEOD, B., CAVIGLIA, A. & PENAGOS, M. (2009). Search engine as a diagnostic tool in difficult immunological and allergologic cases: is google useful? *Internal Medicine Journal*, **39**, 459–464. [58](#)

REFERENCES

- MAEDCHE, A. & STAAB, S. (2004). *Ontology learning*. Springer. 24
- MAILMAN, M.D., FEOLO, M., JIN, Y., KIMURA, M., TRYKA, K., BAGOUTDINOV, R., HAO, L., KIANG, A., PASCHALL, J., PHAN, L. *et al.* (2007). The ncbi dbgap database of genotypes and phenotypes. *Nature Genetics*, **39**, 1181–1186. 69
- MEI, Q., ZHOU, D. & CHURCH, K. (2008). Query suggestion using hitting time. In *Proceedings of the 17th ACM conference on Information and knowledge management*, 469–478, ACM. 7, 29
- METZLER, D. & CROFT, W.B. (2005). A markov random field model for term dependencies. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, 472–479, ACM. 20, 36
- METZLER, D. & KANUNGO, T. (2008). Machine learned sentence selection strategies for query-biased summarization. In *SIGIR Learning to Rank Workshop*, 40–47. 21, 36
- MIKOLOV, T., CHEN, K., CORRADO, G. & DEAN, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*. 11, 13
- MIKOLOV, T., SUTSKEVER, I., CHEN, K., CORRADO, G.S. & DEAN, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119. 11, 13, 30, 35
- MILLER, R.A. (1994). Medical diagnostic decision support systems-past, present, and future a threaded bibliography and brief commentary. *Journal of the American Medical Informatics Association*, **1**, 8–27. 59
- MORENO, J.G. & DIAS, G. (2013). Hultech at the ntcir-10 intent-2 task: Discovering user intents through search results clustering. In *Proceedings of the 10th NTCIR Conference on Evaluation of Information Access Technologies*, NTCIR. 9, 31

REFERENCES

- MORENO, J.G. & DIAS, G. (2016). Search intent mining by word vectors clustering at ntcir-imine. In *Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies*, 41–45, NTCIR. [9](#), [31](#)
- MORENO, J.G., DIAS, G. & CLEUZIQU, G. (2014). Query log driven web search results clustering. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, 777–786, ACM. [8](#), [30](#), [45](#), [50](#), [51](#), [53](#), [54](#), [55](#), [86](#)
- NAVLAKHA, S. & KINGSFORD, C. (2010). The power of protein interaction networks for associating genes with diseases. *Bioinformatics*, [26](#), 1057–1063. [58](#)
- NEWMAN, M.E. (2003). The structure and function of complex networks. *SIAM Review*, [45](#), 167–256. [64](#)
- NGUYEN, B.V. & KAN, M.Y. (2007). Functional faceted web query analysis. In *WWW2007: 16th International World Wide Web Conference*. [7](#), [29](#)
- NGUYEN, T.N. & KANHABUA, N. (2014). Leveraging dynamic query subtopics for time-aware search result diversification. In *Advances in Information Retrieval*, 222–234, Springer. [2](#), [27](#)
- OSIŃSKI, S., STEFANOWSKI, J. & WEISS, D. (2004). Lingo: Search results clustering algorithm based on singular value decomposition. In *Intelligent information processing and web mining*, 359–368, Springer. [42](#)
- RADIVOJAC, P., PENG, K., CLARK, W., PETERS, B., MOHAN, A., BOYLE, S. & MOONEY, S. (2008). An integrated approach to inferring gene–disease associations in humans. *Proteins: Structure, Function, and Bioinformatics*, [72](#), 1030–1037. [58](#)
- RADLINSKI, F., SZUMMER, M. & CRASWELL, N. (2010). Inferring query intent from reformulations and clicks. In *Proceedings of the 19th international conference on World wide web*, 1171–1172, ACM. [8](#), [29](#)

REFERENCES

- REN, P., CHEN, Z., MA, J., WANG, S., ZHANG, Z. & REN, Z. (2015). Mining and ranking users intents behind queries. *Information Retrieval Journal*, **18**, 504–529. [2](#), [27](#)
- ROBERTSON, S. & ZARAGOZA, H. (2009). *The probabilistic relevance framework: BM25 and beyond*. Now Publishers Inc. [18](#), [36](#)
- ROBINSON, P. & MUNDLOS, S. (2010). The human phenotype ontology. *Clinical genetics*, **77**, 525–534. [60](#), [65](#)
- RONG, X. (2014). word2vec parameter learning explained. *arXiv preprint arXiv:1411.2738*. [14](#)
- SAKAI, T. (2005). The reliability of metrics based on graded relevance. In *Information Retrieval Technology*, 1–16, Springer, Jeju Island, Korea. [78](#)
- SAKAI, T. (2011). Ntcireval: A generic toolkit for information access evaluation. In *Proceedings of the Forum on Information Technology*, 23–30. [45](#)
- SAKAI, T. (2014). Statistical reform in information retrieval? *ACM SIGIR Forum*, **48**, 3–12. [45](#)
- SAKAI, T., DOU, Z., YAMAMOTO, T., LIU, Y., ZHANG, M., KATO, M.P., SONG, R. & IWATA, M. (2013a). Summary of the ntcir-10 intent-2 task: Subtopic mining and search result diversification. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, 761–764, ACM. [3](#), [27](#), [28](#), [45](#), [46](#)
- SAKAI, T., DOU, Z., YAMAMOTO, T., LIU, Y., ZHANG, M., SONG, R., KATO, M. & IWATA, M. (2013b). Overview of the ntcir-10 intent-2 task. In *Proceedings of the 10th NTCIR Conference on Evaluation of Information Access Technologies*, NTCIR. [28](#), [43](#)
- SALTON, G. & BUCKLEY, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management*, **24**, 513–523. [36](#)

REFERENCES

- SALTON, G., FOX, E.A. & WU, H. (1983). Extended boolean information retrieval. *Commun. ACM*, **26**, 1022–1036. [16](#)
- SANTOS, R.L., MACDONALD, C. & OUNIS, I. (2010a). Exploiting query reformulations for web search result diversification. In *Proceedings of the 19th international conference on World wide web*, 881–890, ACM. [3](#), [8](#), [9](#), [27](#), [29](#), [33](#)
- SANTOS, R.L., MCCREADIE, R.M., MACDONALD, C. & OUNIS, I. (2010b). University of glasgow at trec 2010: Experiments with terrier in blog and web tracks. In *TREC*. [18](#)
- SANTOS, R.L., MACDONALD, C. & OUNIS, I. (2015). Search result diversification. *Foundations and Trends in Information Retrieval*, **9**, 1–90. [55](#)
- SHI, X. & YANG, C.C. (2006). Mining related queries from search engine query logs. In *Proceedings of the 15th international conference on World Wide Web*, 943–944, ACM. [21](#)
- SOCHER, R., BAUER, J., MANNING, C.D. & NG, A.Y. (2013). Parsing with compositional vector grammars. In *In Proceedings of the ACL conference*, Citeseer. [21](#)
- SONG, F. & CROFT, W.B. (1999). A general language model for information retrieval. In *Proceedings of the eighth international conference on Information and knowledge management*, 316–321, ACM. [19](#)
- SONG, R., LUO, Z., NIE, J.Y., YU, Y. & HON, H.W. (2009). Identification of ambiguous queries in web search. *Information Processing & Management*, **45**, 216–229. [2](#), [7](#), [27](#), [29](#)
- SPARCK JONES, K., WALKER, S. & ROBERTSON, S. (2000). A probabilistic model of information retrieval: development and comparative experiments:: Part 2. *Information Processing & Management*, **36**, 809–840. [18](#)
- SPÄRCK-JONES, K., ROBERTSON, S.E. & SANDERSON, M. (2007). Ambiguous requests: implications for retrieval tests, systems and theories. *ACM SIGIR Forum*, **41**, 8–17. [2](#), [7](#), [27](#), [29](#)

REFERENCES

- STROHMAN, T., METZLER, D., TURTLE, H. & CROFT, W.B. (2005). Indri: A language model-based search engine for complex queries. In *Proceedings of the International Conference on Intelligent Analysis*, 2–6, Citeseer. [4](#), [44](#)
- STUDER, R., BENJAMINS, V.R. & FENSEL, D. (1998). Knowledge engineering: principles and methods. *Data & knowledge engineering*, **25**, 161–197. [24](#)
- SUN, J., QU, H., CHAKRABARTI, D. & FALOUTSOS, C. (2005). Neighborhood formation and anomaly detection in bipartite graphs. In *Data Mining, Fifth IEEE International Conference on*, 8–pp, IEEE. [10](#)
- SUN, P., GAO, L. & HAN, S. (2011). Prediction of human disease-related gene clusters by clustering analysis. *International Journal of Biological Sciences*, **7**, 61. [61](#)
- TANG, H. & NG, J.H.K. (2006). Googling for a diagnosis-use of google as a diagnostic aid: internet based study. *BMJ*, **333**, 1143–1145. [58](#)
- TSAFNAT, G., JASCH, D., MISRA, A., CHOONG, M.K., LIN, F.P.Y. & COIERA, E. (2014). Gene–disease association with literature based enrichment. *Journal of Biomedical Informatics*, **49**, 221–226. [58](#)
- ULLAH, M. & AONO, M. (2016). A bipartite graph-based ranking approach to query subtopics diversification focused on word embedding features. *IEICE TRANSACTIONS on Information and Systems*, **99-D**, 3090–3100. [4](#), [5](#)
- ULLAH, M., AONO, M. & SEDDIQUI, M. (2013a). Estimating a ranked list of human hereditary diseases for clinical phenotypes by using weighted bipartite network. In *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE*, 3475–3478, IEEE, IEEE, osaka, japan. [5](#)
- ULLAH, M.Z., AONO, M. & SEDDIQUI, M.H. (2013b). Sem12 at the ntcir-10 intent-2 english subtopic mining subtask. In *Proceedings of the 10th NTCIR Conference on Evaluation of Information Access Technologies*, NTCIR. [9](#), [31](#)

REFERENCES

- ULLAH, M.Z., AONO, M. & SEDDIQUI, M.H. (2015). Estimating a ranked list of human genetic diseases by associating phenotype-gene with gene-disease bipartite graphs. *ACM Trans. Intell. Syst. Technol.*, **6**, 56:1–56:21. [5](#), [10](#), [38](#)
- ULLAH, M.Z., SHAJALAL, M. & AONO, M. (2016a). Kdeim at ntcir-12 imine-2 search intent mining task: Query understanding through diversified ranking of subtopics. In *Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies*, 60–63, NTCIR. [9](#), [31](#)
- ULLAH, M.Z., SHAJALAL, M., CHY, A.N. & AONO, M. (2016b). Query subtopic mining exploiting word embedding for search result diversification. In *Information Retrieval Technology: 12th Asia Information Retrieval Societies Conference, AIRS 2016, Beijing, China, November 30 – December 2, 2016, Proceedings*, 308–314, Springer International Publishing, Cham. [4](#), [5](#)
- WANG, C.J., LIN, Y.W., TSAI, M.F. & CHEN, H.H. (2013a). Mining subtopics from different aspects for diversifying search results. *Information retrieval*, **16**, 452–483. [2](#), [3](#), [8](#), [27](#), [29](#)
- WANG, J., TANG, G., XIA, Y., ZHOU, Q., ZHENG, F., HU, Q., NA, S. & HUANG, Y. (2013b). Understanding the query: Thcib and thuis at ntcir-10 intent task. In *Proceedings of the 10th NTCIR Conference on Evaluation of Information Access Technologies*, NTCIR. [9](#), [31](#)
- WANG, Q., QIAN, Y., SONG, R., DOU, Z., ZHANG, F., SAKAI, T. & ZHENG, Q. (2013c). Mining subtopics from text fragments for a web query. *Information retrieval*, **16**, 484–503. [3](#), [8](#), [27](#), [30](#)
- WANG, Y., TONG, Y. & ZENG, M. (2013d). Ranking scientific articles by exploiting citations, authors, journals, and time information. In *Twenty-Seventh AAAI Conference on Artificial Intelligence*. [10](#)
- WU, Y., WU, W., LI, Z. & ZHOU, M. (2015). Mining query subtopics from questions in community question answering. In *AAAI*, 339–345. [29](#)

REFERENCES

- XIA, L., XU, J., LAN, Y., GUO, J. & CHENG, X. (2016). Modeling document novelty with neural tensor network for search result diversification. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '16*, 395–404, ACM, New York, NY, USA. [10](#)
- XUE, Y., CHEN, F., DAMIEN, A., LUO, C., LI, X., HUO, S., ZHANG, M., LIU, Y. & MA, S. (2013). Thuir at ntcir-10 intent-2 task. In *Proceedings of the 10th NTCIR Conference on Evaluation of Information Access Technologies, NTCIR*. [9](#), [31](#)
- YAMAMOTO, T., LIU, Y., ZHANG, M., DOU, Z., ZHOU, K., MARKOV, L., KATO, M., OHSHIMA, H. & FUJITA, S. (2016). Overview of the ntcir-12 imine-2 task. In *Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies, NTCIR*. [28](#), [43](#), [45](#), [46](#)
- YANAI, K. *et al.* (2015). Visualtextualrank: An extension of visualrank to large-scale video shot extraction exploiting tag co-occurrence. *IEICE TRANSACTIONS on Information and Systems*, **98**, 166–172. [10](#), [38](#)
- YANG, P., LI, X., WU, M., KWOH, C. & NG, S. (2011). Inferring gene-phenotype associations via global protein complex network propagation. *PloS One*, **6**, e21502. [58](#)
- YILDIRIM, M.A., GOH, K.I., CUSICK, M.E., BARABÁSI, A.L. & VIDAL, M. (2007). Drug-target network. *Nature Biotechnology*, **25**, 1119–1126. [11](#), [60](#)
- YU, H.T. & REN, F. (2014). Search result diversification via filling up multiple knapsacks. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, 609–618, ACM. [10](#)
- YUE, M., DOU, Z., HU, S., LI, J., WANG, X. & WEN, J.R. (2016). Rucir at ntcir-12 imine-2 task. In *Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies*, 36–40, NTCIR. [9](#), [31](#)

REFERENCES

- YUE, Y. & JOACHIMS, T. (2008). Predicting diverse subsets using structural svms. In *Proceedings of the 25th international conference on Machine learning*, 1224–1231, ACM. [10](#)
- ZAMANI, H. & CROFT, W.B. (2016). Estimating embedding vectors for queries. In *Proceedings of the 2016 ACM on International Conference on the Theory of Information Retrieval, ICTIR '16*, 123–132, ACM, New York, NY, USA. [11](#)
- ZHAI, C. & LAFFERTY, J. (2001). A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, 334–342, ACM. [19](#), [36](#)
- ZHANG, Y., DE, S., GARNER, J.R., SMITH, K., WANG, S.A. & BECKER, K.G. (2010). Systematic analysis, comparison, and integration of disease based human genetic association data and mouse genetic phenotypic information. *BMC Medical Genomics*, **3**, 1. [69](#)
- ZHANG, Z. & NASRAOUI, O. (2006). Mining search engine query logs for query recommendation. In *Proceedings of the 15th international conference on World Wide Web*, 1039–1040. [7](#), [29](#)
- ZHENG, G. & CALLAN, J. (2015). Learning to reweight terms with distributed representations. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 575–584, ACM. [11](#)
- ZHU, Y., LAN, Y., GUO, J., CHENG, X. & NIU, S. (2014). Learning for search result diversification. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, 293–302, ACM. [10](#)
- ZOU, H. & HASTIE, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **67**, 301–320. [37](#)