35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

Estimating a Ranked List of Human Hereditary Diseases for Clinical Phenotypes by Using Weighted Bipartite Network

Md. Zia Ullah[1], Masaki Aono[2] and Md. Hanif Seddiqui[3]

***Abstract*— With the availability of the huge medical knowledge data on the Internet such as the human disease network, protein-protein interaction ($\mathcal{PPI}$) network, and phenotype-gene, gene-disease bipartite networks, it becomes practical to help doctors by suggesting plausible hereditary diseases for a set of clinical phenotypes. However, identifying candidate diseases that best explain a set of clinical phenotypes by considering various heterogeneous networks is still a challenging task. In this paper, we propose a new method for estimating a ranked list of plausible diseases by associating phenotype-gene with gene-disease bipartite networks. Our approach is to count the frequency of all the paths from a phenotype to a disease through their associated causative genes, and link the phenotype to the disease with path frequency in a new phenotype-disease bipartite ($\mathcal{PDB}$) network. After that, we generate the candidate weights for the edges of phenotypes with diseases in $\mathcal{PDB}$ network. We evaluate our proposed method in terms of Normalized Discounted Cumulative Gain (*NDCG*), and demonstrate that we outperform the previously known disease ranking method called *Phenomizer*.**

I. INTRODUCTION

One of the formidable tasks in bioinformatics research is to understand the underlying mechanisms of human disease. There are some genes that are responsible for causing human diseases, called disease causative genes or causative genes [1]. 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 [2] and, in turn, causative genes have associations with human diseases [3] as well. Therefore, there might be paths from a phenotype to human hereditary diseases through causative genes with weighting factors along with the edges. Human diseases might be developed through the phenotypical changes due to some causative genes [4,5], and physicians diagnose diseases utilizing their human knowledge of varieties of cases. Wrong selection of clinical features or medical cases may act human severely. Consequently, making the correct diagnosis is questionably the most significant role of the physician. However, disease retrieval system may support physician in diagnosis or treatment practice. In a complex or even in an unknown case of diseases, physicians may get assistance to take decision quickly and efficiently. Therefore, disease retrieval from a set of clinical features is an important and supportive tool for physicians.

The rest of the paper is organized as follows: **Section 2** describes the state of the art while general terminology is articulated in **Section 3**. We introduce our approach in **Section 4**. **Section 5** includes evaluation and experiment. Concluding remarks and some future directions of our work is described in **Section 6**.

II. STATE OF THE ART

The enormous cost of health care is quickly becoming uncontrolled. Over the last decade, there are a number of systems to address the crisis. Moreover, most of the systems are designed to make a prediction about a specific disease or a class of diseases.

Phenomizer is a web-based system that produces a ranked list of hereditary diseases for a set of clinical features [6]. This system only measures the structural similarity of phenotypes between query and diseases using Human-Phenotype-Ontology($\mathcal{HPO}$) [7] by developing a statistical model to assign p values to the resulting similarity scores, which can be used to rank the candidate diseases. However, without considering genetic loci, phenotypic similarity does not alway confirm the relevant plausible diseases.

Another system known as *CARE*, which uses collaborative filtering methods to predict each patient's disease risks based only on their own medical history and that of similar patient's [8]. Moreover, there are some causative genes that can active in the organism in different age of onset.

In the postgenomic era, it is widely established in bioinformatics and system biology to represent associations between biomedical entities as networks and to analyze their topology to get a global understanding of underlying relationships [4]. By means of functional annotation analysis of gene-disease association database, it is indicated a shared genetic origin of human diseases and shown that for most diseases, including mendelian, complex and environmental diseases, functional modules exist [9]. Another study to identify gene-phenotype relationship instead of finding the gene-disease relationship directly states that similar phenotypes are caused by functionally related genes [10].

III. GENERAL TERMINOLOGY

This section introduces some basic definitions of terminology to familiarize the readers with the notions used throughout the paper. It includes the definitions of phenotype-gene, and gene-disease bipartite networks.

[1]Md. Zia Ullah is with Masters Student of the Department of Computer Science and Engineering, Toyohashi University of Technology, 1-1 Hibarigaoka, Tempaku-Cho, Toyohashi, 441-8580, Aichi, Japan, arif@kde.cs.tut.ac.jp

[2]Masaki Aono is with the Professor of the Department of Computer Science and Engineering, Toyohashi University of Technology, 1-1 Hibarigaoka, Tempaku-Cho, Toyohashi, 441-8580, Aichi, Japan, aono@tut.jp

[3]Md. Hanif Seddiqui is with Associate Professor of the Department of Computer Science and Engineering, University of Chittagong, Hathazari-4331, Chittagong, Bangladesh, hanif@cu.ac.bd

978-1-4577-0216-7/13/$26.00 ©2013 IEEE

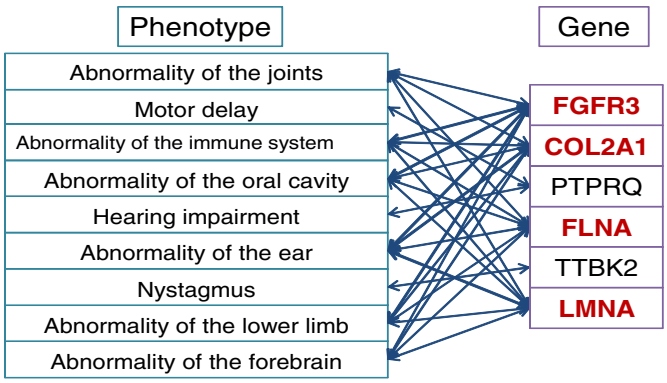

Fig. 1. Phenotype to Gene Bipartite Network with unit Edge Weight

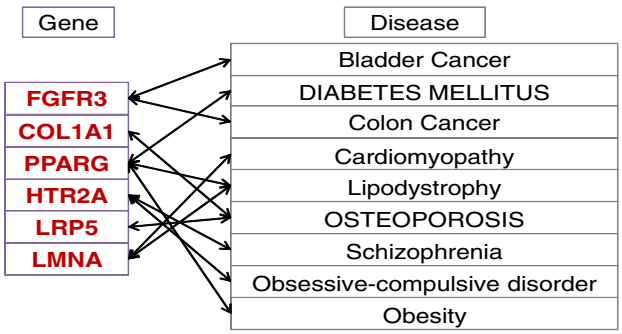

Fig. 2. Gene to Disease Bipartite Network with unit Edge Weight

A. *Phenotype-Gene bipartite network*

The correspondence between gene and phenotype is a many-to-many relation in which any given gene corresponds to multiple different phenotypes and there are different genes corresponding to a given phenotype [11]. For example, 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). The associations of phenotypes with genes are represented as a bipartite network (Phenotype-Gene Bipartite ($\mathcal{PGB}$) network), all the edges between phenotypes and genes are initialized to one as shown in Fig. 1. All the genes in bold letters in Fig. 1 are known as disease causative genes.

B. *Gene-Disease bipartite network*

The gene-disease network consists of two types of nodes (gene and disease) [4]. Gene and disease nodes are connected through edges if the corresponding gene-disease association is covered in the gene-disease database. A set of disease causative genes FGFR3 (2261), COL2A1 (1280), FLNA (2316), and HTR2A (3356) are associated with a set of human diseases Bladder Cancer (OMIM:109800), Colon Cancer (OMIM:114500), Wagner syndrome (143200), SCHIZOPHRENIA (OMIM:181500), and Heterotopia (OMIM:300049). The associations are depicted in Fig. 2 as a bipartite network (Gene-Disease Bipartite ($\mathcal{GDB}$) network).

IV. Our Approach

In our methodology, firstly, the gene-disease bipartite ($\mathcal{GDB}$) network is extended by using $\mathcal{PPI}$ network and known gene-disease associations. Secondly, all hidden paths of phenotypes with diseases are explored by associating phenotype-gene ($\mathcal{PGB}$) with extended gene-disease bipartite ($\mathcal{EGDB}$) networks considering the common causative genes in both networks. Finally, the phenotype-disease associations are represented as a bipartite (Phenotype-Disease Bipartite ($\mathcal{PDB}$)) network, and is weighted using our proposed model Bidirectional Phenotype-disease Weight ($\mathcal{BPW}$) method [12]. Using the weighted $\mathcal{PDB}$ network, we produce a ranked list of candidate diseases for a set of clinical phenotypes. For evaluating our system, we make a set of queries from some specific branch of Human-Phenotype-Ontology ($\mathcal{HPO}$).

A. *Extension of GDB network*

There is more and more evidence that most human diseases cannot be attributed to single gene but arise due to complex interactions between multiple genetic variants and environmental risk factors [5]. Since disease causative genes which are more likely to interact with each other through their protein products, exploiting protein-protein interactions ($\mathcal{PPI}$) can greatly increase the likelihood of finding positional candidate disease genes [10]. We may consider the first-neighboring genes of causative genes in $\mathcal{PPI}$ network as susceptible to diseases. Through this idea, we explore more candidate causative genes, and ultimately extend $\mathcal{GDB}$ network. The complete procedure is outlined in Algorithm 1. It requires three basic operations that are applied to $\mathcal{PPI}$, $\mathcal{GDB}$, and $\mathcal{EGDB}$ networks. The first operation called *getCausativeGene* returns the set of causative genes ($\mathcal{CG}$) of a disease d from $\mathcal{GDB}$ network. The second operation denoted *getFirstNeighbor* returns the set of first-neighboring genes ($\mathcal{NG}$) of a causative gene $cg \in \mathcal{CG}$ from $\mathcal{PPI}$ network. Finally, third operation updates the $\mathcal{EGDB}$ by adding new edge between each candidate gene $ng \in \mathcal{NG}$ and disease d along with the existing edges of $\mathcal{GDB}$ network.

B. *Association of Phenotype with Disease*

It is strongly believed that a phenotype is associated with a set of genes, and in turn, a causative gene is also associated with a set of diseases. Therefore, there might be a path from a phenotype to a disease through a gene. We link a phenotype with a disease by searching all paths from a phenotype to one or more genes, and in turn, genes to a disease. We represent the associations of phenotypes with diseases as a bipartite ($\mathcal{PDB}$) network where each edge is labelled with a frequency. The frequency of an edge is the total number of distinct paths from a phenotype to a disease through one or more genes.

Algorithm 1: ExtendCausativeGene($\mathcal{PPI}$, $\mathcal{GDB}$)
A naïve algorithm for exploring disease causative genes

Input: $\mathcal{PPI}$ and $\mathcal{GDB}$ Network
Output: $\mathcal{EGDB}$ Network

$\mathcal{D} \leftarrow$*getAllDisease*($\mathcal{GDB}$)
$\mathcal{EGDB} \leftarrow \emptyset$
for *disease* $d_i \in \mathcal{D}$ **do**
  $\mathcal{CG} \leftarrow$ *getCausativeGene*($\mathcal{GDB}, d_i$)
  for *causative gene* $cg_j \in \mathcal{CG}$ **do**
    $\mathcal{EGDB} \leftarrow \mathcal{EGDB} \cup \{cg_j, d_i\}$
    $\mathcal{NG} \leftarrow$ *getFirstNeighbor*($\mathcal{PPI}, cg_j$)
    for *gene* $ng_k \in \mathcal{NG}$ **do**
      $\mathcal{EGDB} \leftarrow \mathcal{EGDB} \cup \{ng_k, d_i\}$
return $\mathcal{EGDB}$

C. Candidate Weight Generation

In this section, we elucidate the ways of weighting $\mathcal{PDB}$ network. The $\mathcal{PDB}$:= ($\mathcal{P}$+$\mathcal{D}$, $\mathcal{E}$) is a bipartite network where $\mathcal{P}$ is the set of phenotypes, $\mathcal{D}$ is the set of diseases, $p_1 \in \mathcal{P}$, $d_1 \in \mathcal{D}$, and $(p_1, d_1) \in \mathcal{E}$ is an edge in $\mathcal{PDB}$. The edge weight is measured by applying *BM25* [13], and our proposed $\mathcal{BPW}$ individually.

1) BM25 Weight: The *BM25* weight of a phenotype p_i on a disease d_j is calculated using the following equation:

$$weight(p_i, d_j) = \frac{\mathcal{TF}_{p_i,d_j} \cdot (k_1 + 1)}{k_1 \cdot ((1-b) + (b \cdot \frac{l_d}{avgl_{\mathcal{D}}})) + \mathcal{TF}_{p_i,d_j}} \times \log \frac{|\mathcal{D}| - |\{d \in \mathcal{D} | p_i \in d\}| + 0.5}{|\{d \in \mathcal{D} | p_i \in d\}| + 0.5} \tag{1}$$

where $\mathcal{TF}_{p_i,d_j}$ is the p_i's term frequency in the disease d_j, l_d is the length of disease d i.e. the number of phenotypes occurring in it, $avgl_{\mathcal{D}}$ is the average disease length, $|\{d \in \mathcal{D} | p_i \in d\}|$ is the total number of diseases where the phenotype p_i occurs.

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 $k_1 = \{1.2, 1.3, 1.4, ..., 2.0\}$ and $b = \{0.50, 0.55, 0.60, ..., 0.90\}$. For each pair of k_1 and b, we produce the average $\mathcal{F}$-measure [14]. 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 1.

2) BPW Weight: The weight using our proposed method $\mathcal{BPW}$ of a phenotype p_i on a disease d_j is calculated as follows:

$$weight(p_i, d_j) = (\frac{avgl_{\mathcal{D}}}{l_{d_j}} \cdot \frac{|p_i \in d_j|}{\Sigma_x |p_x \in d_j|}) + (\frac{avgl_{\mathcal{P}}}{l_{p_i}} \cdot \frac{|p_i \in d_j|}{\Sigma_y |(p_i \in d_y) : d_y \in \mathcal{D}|}) \tag{2}$$

where l_{d_j} is the length of disease d_j, $avgl_{\mathcal{D}}$ is the average disease length, $|p_i \in d_j|$ is the number of times the phenotype p_i occurs in the disease d_j, $\Sigma_x |p_x \in d_j|$ is the number of times of all the phenotypes p_x occurs in the disease d_j, l_{p_i} is the length of phenotype p_i i.e. the number of diseases where the phenotype p_i appears, $avgl_{\mathcal{P}}$ is the average phenotype length, $\Sigma_y |(p_i \in d_y) : d_y \in \mathcal{D}|$ is the count of all the diseases d_y where the phenotype p_i occurs.

The $\mathcal{BPW}$ weight is the sum of the importance weight of phenotype p_i for disease d_j, and disease d_j for phenotype p_i. This is bidirectional importance from both phenotype and disease respectively. The procedure to generate the candidate weight between every edge of the $\mathcal{PDB}$ network is outlined in Algorithm 2. It requires three basic operations. The first operation is *getFirstNeighbor* that returns a list of diseases which are associated with a phenotype. The second operation is to apply a specific weight equation i.e. equation (1) or (2), between a phenotype with its first-neighboring disease to estimate the candidate weight. Third operation updates the weighted phenotype-disease bipartite ($\mathcal{WPDB}$) network by adding the weighted edge.

Algorithm 2: CandidateWeightGeneration($\mathcal{PDB}$, *Weight Equations*)
An algorithm for generating edge weight of the phenotype-disease bipartite ($\mathcal{PDB}$) network

Input: $\mathcal{PDB}$) Network, Weight Equations
Output: $\mathcal{WPDB}$ Network
$\mathcal{WPDB} \leftarrow$ *empty*
for *phenotype* $p_i \in \mathcal{PDB}$ **do**
  $\mathcal{D} \leftarrow$ getFirstNeighbor($\mathcal{PDB}$,p_i)
  for *disease* $d_j \in \mathcal{D}$ **do**
    $weight \leftarrow$ Apply a specific weight equation between p_i and d_j
    $\mathcal{WPDB} \leftarrow \mathcal{WPDB} \cup \{(p_i, d_j), weight\}$
return $\mathcal{WPDB}$

D. Probable Disease Retrieval System

In this section, we describe the method of retrieving ranked list of candidate diseases for a set of clinical phenotypes. Our system uses the $\mathcal{WPDB}$ network for predicting plausible diseases for a given set of phenotypes. Let us assume that a given phenotype set is $\mathcal{Q} = \{p_1, p_2, ..., p_k\}$. The specificity of a disease, d to the given phenotypes set, $\mathcal{Q}$ is defined as follows:

$$\Phi_d = \sum_{p_{i=1}}^{|\mathcal{Q}|} w(p_i, d) \;\; if (p_i, d) \in \mathcal{WPDB} \tag{3}$$

Using equation 3, we may have a set of diseases with their weights. The cumulative weight is calculated for every distinct disease. Then, the diseases are sorted in descending order according to their cumulative weights. Now, physician can observe the top-5 or top-10 diseases in the ranked list to diagnose more precisely with the help of our assistive disease retrieval system.

V. Experiments and Evaluation

We selected $\mathcal{PGB}$, $\mathcal{HPO}$, and disease-phenotype annotation from [7] website, and $\mathcal{PPI}$, $\mathcal{GDB}$ from [4] website. In $\mathcal{PGB}$ network, there are *6,327* phenotypes and *1,807* genes. $\mathcal{GDB}$ network consists of *1,271* causative genes and *1,540* diseases, PPI includes *951* genes. After applying Algorithm 1, we found *1,446* causative genes in the $\mathcal{EGDB}$ network.

We choose to evaluate using Normalized Discounted Cumulative Gain (*NDCG@k*) [14], which credits system with high precision at top-k ranks. Let $g_1, g_2, \cdots, g_k$ be the gain values associated with the top-k diseases, where g_i is the gain value for relevance grade ξ at rank i. Then, *NDCG* value is defined as follows:

$$NDCG_k = \frac{DCG_k}{IDCG_k} \text{ where } DCG_k = \sum_{i=1}^{k} \frac{2^{g_i}-1}{lg(i+1)}$$

and $IDCG_k$ denotes the DCG_k value for an ideal ranked list. For estimating *NDCG@22*, we make a set of *42* queries $\mathcal{Q}$, which are chosen from some specific branch of $\mathcal{HPO}$ e.g. "Abnormality of abdomen, and Abnormality of immune system". Then, a ranked list of candidate diseases is retrieved for every query based on *BM25*, and $\mathcal{BPW}$ individually. Every disease in the ranked list is elucidated for its relevance to the query phenotypes. Relevance grade is measured on a 5-level scale i.e irrelevant, marginally relevant, partially relevant, fairly relevant, and highly relevant using the Jaccard's Index of query and disease annotated phenotypes.

NDCG@22 is further measured for the ranked list of diseases, which is produced by *Phenomizer*. *Phenomizer* does not produce any *NDCG* measure, however, we use the same set of queries $\mathcal{Q}$ as *BM25* and $\mathcal{BPW}$. The comparison result of our system with *Phenomizer* and *BM25* for *NDCG@22* is depicted in Fig. 3. It is clearly turned out that our system outperforms *Phenomizer*, and is to some extent better than *BM25*. We does not compare our system with other state of the art works e.g. POSSUM, The London Dysmorphology Database (LDDB), as well as the search routine available with the OMIM, and Orphanet. The reason is that, these systems do not provide explicit rankings or measures of plausibility for the potential long lists of candidate diseases.

VI. Conclusion and Future Direction

In this paper, we proposed a new method for estimating a ranked list of plausible diseases. Our approach was to explore all the paths from a phenotype to a disease through causative genes, and link the phenotype to the disease with path frequency. After that, we generated the candidate weights for the edges of phenotypes with diseases in $\mathcal{PDB}$ network. We experimented with and evaluated our system by measuring *NDCG@22*. It is clearly shown that we outperform the previously known disease ranking method called *Phenomizer*.

One of our future targets is to extend gene-disease bipartite network by exploring more causative genes using *PPI*, and gene expression. Along with genetic linkage, we will further implement the structural similarity of disease annotated phenotypes and query phenotypes using *HPO* for refining the ranked list of candidate diseases for differential diagnosis.

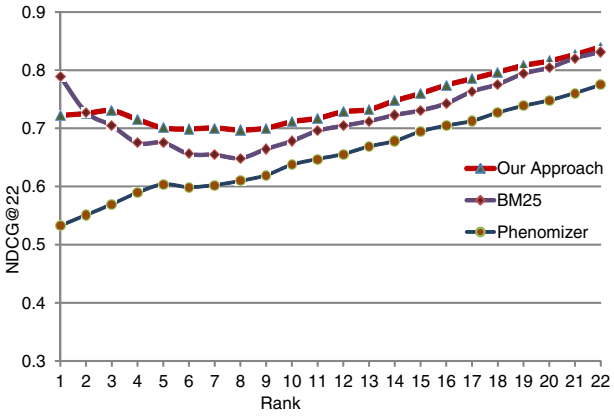

Fig. 3. Comparison of Our Approach with Phenomizer and BM25 for NDCG@22

VII. Acknowledgments

This research was partially supported by the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Grant-in-Aid(C) 23500119.

References

[1] F. Barrenäs, S. Chavali, A. Alves, L. Coin, M. Jarvelin, R. Jörnsten, M. Langston, A. Ramasamy, G. Rogers, H. Wang *et al.*, "Highly interconnected genes in disease-specific networks are enriched for disease-associated polymorphisms," *Genome Biology*, vol. 13, no. 6, p. R46, 2012.

[2] P. Yang, X. Li, M. Wu, C. Kwoh, and S. Ng, "Inferring gene-phenotype associations via global protein complex network propagation," *PloS one*, vol. 6, no. 7, p. e21502, 2011.

[3] T. Hwang, G. Atluri, M. Xie, S. Dey, C. Hong, V. Kumar, and R. Kuang, "Co-clustering phenome–genome for phenotype classification and disease gene discovery," *Nucleic Acids Research*, 2012.

[4] K. Goh, M. Cusick, D. Valle, B. Childs, M. Vidal, and A. Barabási, "The human disease network," *Proceedings of the National Academy of Sciences*, vol. 104, no. 21, p. 8685, 2007.

[5] J. Hardy and A. Singleton, "Genomewide association studies and human disease," *New England Journal of Medicine*, vol. 360, no. 17, pp. 1759–1768, 2009.

[6] S. Köhler, M. Schulz, P. Krawitz, S. Bauer, S. Dölken, C. Ott, S. Mundlos, and P. Robinson, "Clinical diagnostics in human genetics with semantic similarity searches in ontologies," *The American Journal of Human Genetics*, vol. 85, no. 4, pp. 457–464, 2009.

[7] P. Robinson and S. Mundlos, "The human phenotype ontology," *Clinical Genetics*, vol. 77, no. 6, pp. 525–534, 2010.

[8] D. Davis, N. Chawla, N. Christakis, and A. Barabási, "Time to CARE: a collaborative engine for practical disease pediction," *Data Mining and Knowledge Discovery*, vol. 20, no. 3, pp. 388–415, 2010.

[9] A. Bauer-Mehren, M. Bundschus, M. Rautschka, M. Mayer, F. Sanz, and L. Furlong, "Gene-disease network analysis reveals functional modules in mendelian, complex and environmental diseases," *PloS one*, vol. 6, no. 6, p. e20284, 2011.

[10] M. Oti, B. Snel, M. Huynen, and H. Brunner, "Predicting disease genes using protein–protein interactions," *Journal of Medical Genetics*, vol. 43, no. 8, pp. 691–698, 2006.

[11] R. Lewontin, "The genotype/phenotype distinction," in *The Stanford Encyclopedia of Philosophy*, 2011st ed., E. N. Zalta, Ed., 2011.

[12] M. Ullah, M. Aono, and M. Seddiqui, "Ranking human genetic diseases by associating phenotype-gene with gene-disease bipartite graphs. unpublished manuscript," *Knowledge and Information Systems*, 2013.

[13] S. Robertson, S. Walker, and M. Beaulieu, "Okapi at trec-7: automatic ad hoc, filtering, vlc and interactive track," *Nist Special Publication SP*, pp. 253–264, 1999.

[14] D. Harman, "Information retrieval evaluation," *Synthesis Lectures on Information Concepts, Retrieval, and Services*, vol. 3, no. 2, pp. 1–119, 2011.