Relevance Search via Bipolar Label Diffusion on Bipartite Graphs

Dr. Zhang Liang[1]

[1]

Received: 13 December 2011 Accepted: 2 January 2012 Published: 15 January 2012

Abstract

The task of relevance search is to find relevant items to some given queries, which can be viewed either as an information retrieval problem or as a semi-supervised learning problem. In order to combine both of their advantages, we develop a new relevance search method using label diffusion on bipartite graphs. And we propose a heat diffusion-based algorithm, namely bipartite label diffusion (BLD). Our method yields encouraging experimental results on a number of relevance search problems.

Index terms— Relevance search, ranking, graph diffusion, bipartite graphs.

1 I. INTRODUCTION

he problem of relevance search (RS) has been recognized as an important and interesting problem in machine learning and information retrieval, which refers to finding relevant items to a small query set given by the user. Along with the ability to find a cluster of data that shares some common properties to the query, it is also a primary goal of the information retrieval (IR) systems. Typical applications of RS include the discovery of relevant words in a text corpus [1][2][3][4] , answers to community questions [5,6] , features in conceptlearning problems [7] , recommendations in collaborative filtering systems [8,9] , among others.

Google Sets [10] is a well-known and successful representative of RS systems that has been widely used. It uses vast amount of web-pages to create a list of related items from a few examples, such as people, movies, words, places, etc. Most of the other RS systems are similar to Google Sets in that they perform some ranking algorithm on a large corpus of documents or web-pages. Ghahramani and Heller [11] proposed the Bayesian Sets algorithm that uses a model-based concept of clusters and performs Bayesian inference to compute the ranking scores of items. Sun et al. [12] used bipartite graphs to model the data and introduced random walks with restarts to rank the items.

The RS problem can be viewed from several angles. First, finding relevant items to the query is a standard IR task [11,13] , which may be solved using classic IR algorithms, such as HITS [14] , nearest neighbors, naïve Bayes and Rocchio's algorithm. These algorithms can compute a list of items ordered by the relevance to the query. However, there is no explicit boundary between "relevant" and "irrelevant" items. Second, RS can be interpreted as a special case of semi-supervised learning (SSL) problem with a few positive examples given as the query [11] .

It is essentially a one-class classification/clustering problem, which can be solved using one-class learning techniques like mappingconvergence [15] and OC-SVMs [16,17] . The relevant and irrelevant items can be explicitly classified through SSL algorithms. However, most of these algorithms do not provide the rank order of items, which is of importance in IR systems.

In this paper, we propose and evaluate a new RS method called bipartite label diffusion (BLD) that can be seen as a hybrid between IR and SSL. BLD is a diffusion-based algorithm that works on bipartite graph. User queries are mapped to vertices with positive labels on the graph. BLD performs local computation at every vertex of the graph and develops a global classifier through the label diffusion process. The diffusion process is often modeled using a Markov chain on the graph. In BLD, we modified the diffusion model by adopting the "label spreading" method [18] to allow negative labels diffuse on the graph. If the word/document is relevant to the queries, the score value is positive, otherwise the value is negative. The more the item is relevant, the higher the score is. Thus BLD can be seen as a semi-supervised learning method.

The rest of this paper is organized as follows. We first model the corpus data as document-word bipartite graphs in section 2. The detail of the BLD algorithm is given in section 3. In section 4, we present empirical results to demonstrate the effectiveness of BLD on relevance search problems. Finally, some concluding remarks are given in section 5.

2 II. BIPARTITE GRAPH MODEL

Bipartite graph models have been successfully applied in many fields such as text clustering [19][20][21] , collaborative filtering [22,23] , and content-based image retrieval [24] . The first step of our method is to model the document-word dataset as a bipartite model. We also employ a query expansion scheme to enhance the searching capacity. Detailed description is given in this section.

Suppose we are given a set of n documents , }i j i j E d w d w = ? ? D

W is the set of edges between two sets of vertices D and W . In the bipartite graph G, there are no edges between words or between documents. An edge e (i, j) exists if word w j appears in document d i . which is denoted by j i w d . And it indicates an association between a word and a document, which may be quantitatively expressed by assigning positive weights on the edges. Assume we are given the user's query set of wordsQ W (, Q Q ? ? ? W W W), or documents Q D (, Q Q ? ? ? D D W)

, or both of them, the RS problem can be cast to a semi-supervised learning task where the query set contains the initial labeled examples. Fig. 1 shows a simple example of the bipartite graph model.

Figure 1: Example of a document-word bipartite graph which has 3 documents and a vocabulary of 4 words. The user's query set is1 { } Q w = W and 1 { } Q d = D

. The goal of this similar sets retrieval problem is to discover a set of words that are similar to Q W , and a set of documents that are similar to Q D There are many edge-weighting schemes in information retrieval research, among them we adopt the popular "tf-idf" weight:ln#(,) tf-idf (,) ln ln #(,) { : ~} i j i j i k j k d w n d w d w d w d = ? ? where #(,)

i j d w is the number of occurrence of word w j in document d i . Words with high tf-idf values imply a strong relationship with the document they appear in.

We can define an n m × edge weight matrix M: { , tf-idf (,), if edge (,) exists, 0, otherwise.i j i j i j d w d w M =

To interpret the edge weight matrix from the perspective of Markov chains on graph, which will be described in section 3, we define the transition probability from d i to w j as:, , ,**1**m i j i j i p p P M M = = ?**(1)**

And the transition probability from w j to d i is given by, , , , 1n j i i j p j p Q M M = = ?**(2)**

The Markov transition matrix P normalizes M such that every row sum up to 1, and Q normalizes M T such that every row sum up to 1.

3 Thus the (

) ( ) m n m n + × + adjacency matrix A of the bipartite graph may be written as:? ? = ? ? ? ? 0 Q A P 0 ,

Where the vertices of the graph is ordered such that the first m vertices index the words and the last n vertices index the documents.

Note that bipartite graph can also be a good model in the scenario of collaborative filtering or recommendation systems. A standard collaborative filtering problem often involves a user set U , an item set I , and a vote set V . In a similar manner, we can build a bipartite graph (, ,)E = G U I

. The normalized vote scores of V can directly serve as the weight matrix. Moreover, the aim of collaborative filtering is to discover new items that an active user may like based on her voted items, which could be easily translated to a RS problem by finding similar items to the user's favorite ones (items with high voting scores).

4 III.

5 LABEL DIFFUSION ON BIPARTITE GRAPHS

In general, there are two primary approaches to cluster vertices on a bipartite graph: one is to partition the graph to disjoint parts corresponding to different clusters [25,26] ; the other is to compute a rank value for each vertex indicating the probability that the vertex is in a cluster [24,27] . To capture the uncertainties on datacluster assignments, we investigate the problem from the latter angle.

6 a) Markov Chains on the Graph

In a bipartite graph, there are no direct relationships among the words or among the documents. However, in diffusion-based methods [18,27] , the strength of the similarities among elements on the same side of the bipartite graph may be captured by a local probability evolution process, which can be integrated to obtain a global view of the data.

We define a Markov chain to describe the diffusion process over the bipartite graph. Assume there is a discrete time random walk on the bipolar graph G, the row-normalized adjacency matrix A is the one-step transition matrix that , i j A is the probability of moving to the jth vertex of v given that the current step is at the ith

vertex. More specifically, P is the document-word transition matrix containing the transition probabilities ofw 1 d 2 d 3 w 2 w 3 w 4 W (D D D D) D**2012**

Year moving from a document vertex to a word vertex, and Q is the word-document transition matrix. Since the document-word graph is bipartite, all the paths from w i to w j must go through vertices in D . As is shown in [15, 18], the similarity of two words w i and w j , is a quantity proportional to the probability of direct transitions between them, denoted by p(w i , w j), (,) ( ) (|) = = ? ?

Obviously, it is a 2-step stochastic process that first maps w i to the document set, and then maps the documents back to w j .

Similarly, the conditional transition probability from d i to d j is given by, , ,**(|)**

7 ? ?

Therefore, by formulating the relationship between documents and words as Markov chains on the bipartite graph, we have the word-word transition probability matrix QP , and the document-document transition probability matrix PQ .

8 b) Diffusion Process

Intuitively, our bipolar diffusion framework works as following: First, we construct a bipartite graph with two poles as is described in section 2.3. The heat pole stands for the words and documents that are most relevant to the query, while the cold pole contains the "strongest negative" words and documents. Then a certain amount of heat is injected to the graph through the heat pole, and the cold polar extracts the heat out of the system as a heat sink. And the heat diffuses through the edges of the graph. Since the system has two poles, we name the heat diffusion as the "bipolar diffusion process".

The diffusion process may be thought of as a Markov chain on the graph. The fundamental property of a Markov chain is the Markov property, which make it possible to predict the future state of a system from its present state ignoring its past states. We denote p **??**t) and q (t) as the labels of documents and words at time t. The Markov process then defines a dynamic system,(1) ( ) ( ) t t t + = ? = Q QP p q p (**3**) (1) ( ) ( ) t t t + =? = P PQ q p q (4)

This simple 2-step diffusion process captures the interaction between the two sets of vertices on the graph. It requires p **??**t) and q (t) be the probability distributions of Markov states with non-negative values.

However, in our bipolar graph diffusion framework, the vertices of the cold pole are labeled by negative values. In the following, we cast the diffusion process as a semisupervised learning problem, which makes it possible to diffuse heat and cold simultaneously on the graph. First, we use an (n+2)-vector p **??**0) to denote the initial labels of documents, ? (0) {1, 0,...0, 1} T n = ? p ; and an (m+2)-vector q (0) to denote the initial labels of words,? (0) {1, 0,...0, 1} T m = ? q

. The virtual vertices of heat pole are labeled by positive values, which allow heat diffuses through the graph; while negative values representing "cold" are assigned to the virtual vertices of cold pole.

And the vertices keep exchanging these two kinds of energy as the diffusion process proceeds.

Then we adopt the "label spreading" method [18] to allow negative labels diffuse on the graph. The iteration in Eq. (3) and Eq. (**??**) can be rewritten as,(1) (0) ( ) (**1**)t t ? ? + = + ? ? Q p p q (1) (0) ( ) (**1**)t t ? ? + = + ? ? P q q p

Where ? and ? are scaling parameters both in (0,1), which specify the relative amount of the heat/cold a vertex received from its neighbors and its initial label information. To simplify the diffusion process, we set 1 2 ? ? = = . The iteration equations indicate that when1 t ? (5) (6)

Where and . Since P ? and Q ? are row-normalized matrices, Eq. (**??**) follows that when t ? ? , (1) t+ p converges to,

And Eq. (**??**) converges to,

9 Global Journal of Computer Science and Technology

Volume XII Issue X Version I21 (D D D D) D (1) (0) ( ) (0) (0)**(1) 1 (0)**(0) (0) 0 1 1 1 1 () 2 2 4 4 11 ( )( ) 24 t t t t it i ?? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? Q Q QP ? Q ? p p q p q p p q p (1)**(0) ( ) (0) (0) (1) 1 (0)**(0) (0) 0 1 1 1 1 () 2 2 4 4 11 ( )( ) 24 t t t t it i ?? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? P P PQ Î?" P Î?" q q p q p q q p q 1 4 ? ? QP ? ? 1 4 ? Î?" PQ ? ? ( ) 1 (0) (0) 11 ( ) ( ) 24 ?? ? ? ? I ? Q p p q ( ) 1 (0)**(0) 11 ( ) ( ) 24 ??**? ? ? I Î?" P q q p IV. EXPERIMENTS

Our proposed BLD method is applicable to discover similar items to a query set from a corpus of text data or user rating data. In this section, we experiment with our model on a set of RS problems.

The experiments are performed on two standard text datasets: Reuters-21578 1 , and a collaborative filtering dataset: MovieLens 2

10 Datasets

. All of the text datasets were preprocessed by removing the stopwords and stemming. And for Reuters-21578, we use a subset of the ten most frequent categories with highest number of positive examples, namely Reuters-10.

The main features of the datasets are summarized in Table **??**. We conducted two experiments on both the text dataset Reuter-10 and the movie rating dataset MovieLens to evaluate the RS ability of our method. The results and comparisons with Google Sets 3 , Internet movie database (IMDB) 4 query: science + technology , and Bayesian Sets are given in tables 2, 3 and 4. Concerning the application of movie recommending, the query takes the form of a set of movies. We regard movies as documents and users as words, and the rating scores are equivalent to word frequencies. Thus our algorithm can be easily adapted to collaborative filtering datasets. Among the tested algorithms and systems, Google Sets is a baseline RS system that is based on vast amount of web data. Although the Google Sets algorithm is not available for us to run it on our datasets, it is still informative and worth to be compared with; IMDB provides movie recommendations by generating a list of 5 movies most related to the query, which is based on collaborative filtering technology; Bayesian Sets views RS as a Bayesian inference problem and gives the corresponding ranking algorithm. We experiment with an online Bayesian Sets recommending system on Movie Lens dataset.

From the query results of the relevant words discovery task and the movie suggestion task, we can make the following comments: 1. Table **??** shows that both Google Sets and our method can achieve to some extent similar sets to the query. There is not an objective standard to tell the exact similarity between words or movies. However, the words that our method retrieved are obviously more sensible than some results of Google Sets, e.g. "war" and "Intel" for the query of "market" and "price". We think this is because Google Sets and our method have different learning mechanisms and are based on different corpuses. 2. Our method serves as a good algorithm for recommending systems on the MovieLens dataset.

The recommended movies to the query "Full Metal Jacket" are all related to war. And for "The Graduate", most of the results are romance and drama movies. We notice that IMDB's suggestions are often popular and new, while Bayesian Sets and our method, limited by the MovieLens dataset, tend to generate classic movies.

11 V. CONCLUSION

In this paper we developed a new graph diffusion algorithm for RS. We used bipartite graphs to model the relationships between documents and words. We also modeled the diffusion process using a Markov chain on the graph, and presented the corresponding label diffusion algorithm. In future work, we will extend the proposed method to other applications (e.g. social networks, question answering systems).

VI. [1]

1 

Figure 1: 1 {

[1]http://www.daviddlewis.com/resources/testcollections/reuters21578/ 2 http://www.grouplens.org/system/files/ml-data.tar___0.gz 3 http://labs.google.com/sets 4 http://www.imdb.com/

I

	# documents/uses	# words/items	# nonzero entries
Reuters-10	9,989	5,180	373,440
MovieLens	943	1,682	100,000

Figure 2: Table I :

II

Sets and Bld Based on the Same Given Queries. Bld

Runs on Reuters-10			
		query: market + price	
Google Sets	BLD	Google Sets	BLD
science	scienc	market	market
technology	technologi	price	pric
business	univers	overview	money
sports	engineer	view	stock
health	educat	risk	valu
entertainment research		gains	bond
education	comput	forecasts	busi
politics	life	war	compani
travel	develop	Intel	trade
computers	cultur	losses	economic

Figure 3: Table II :

III

Query: Full Metal Jacket (1987)

Google Sets	IMDB	Bayesian Sets	BLD
Saving Private Ryan (1998)	Platoon (1986)	The Terminator (1984)	Platoo
Apocalypse Now (1979)	All Quiet on the Western Front (1930)	Star Wars Episode V (1980)	Rambo Blood
Platoon (1986)	Cidade de Deus (2002)	Raiders of the Lost Ark (1981)	Braveh
Pulp Fiction (1994)	Batoru Rowaiaru (2000)	Aliens (1986)	Apocal (1979)
Hamburger Hill (1987)	If? (1968)	Die Hard (1988)	Star Episod

Figure 4: Table III :

IV

Query: The Graduate (1967)			
Google Sets	IMDB	Bayesian Sets	BLD
Chinatown (1974)	Mysterious Skin (2004)	Casablanca (1942)	Casablanca (1942)
Midnight Cowboy (1969)	Giant (1956)	The Wizard of Oz (1939)	Annie Hall (1977)
Annie Hall (1977)	The Notebook (2004)	One over Nest (1975) Flew the Cuckoo's	Gone with the Wind (1939)
Taxi Driver (1976)	Bigfish (2003)	The Godfather (1972)	To Mockingbird a Kill (1962)
Bonnie and Clyde (1967)	Notes on a Scandal (2006)	Amadeus (1984)	Giant (1956)

Figure 5: Table IV :

.1 ACKNOWLEDGMENT

[Google and Sets] , Google Google , Sets . <http://labs.google.com/sets>

[Deng et al. ()] 'A generalized Co-HITS algorithm and its application to bipartite graphs'. H Deng , M R Lyu , I King . *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, (the 15th ACM SIGKDD international conference on Knowledge discovery and data mining) 2009. p. .

[Lavrenko ()] *A generative theory of relevance*, V Lavrenko . 2004. University of Massachusetts Amherst

[Blondel et al. ()] 'A measure of similarity between graph vertices: Applications to synonym extraction and web searching'. V D Blondel , A Gajardo , M Heymans , P Senellart , P V Dooren . *Siam Review* 2004. 46 (4) p. .

[Kleinberg ()] 'Authoritative sources in a hyperlinked environment'. J M Kleinberg . *J. ACM* 1999. 46 (5) p. .

[Lin ()] 'Automatic retrieval and clustering of similar words'. D Lin . *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING/ACL-98*, (the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING/ACL-98) 1998. p. .

[Cohen ()] 'Automatically extracting features for concept learning from the web'. W W Cohen . *Proceedings of the 17th International Conference on Machine Learning*, (the 17th International Conference on Machine Learning) 2000. p. .

[Ghahramani and Heller ()] 'Bayesian sets'. Z Ghahramani , K A Heller . *Advances in Neural Information Processing Systems*, 2005. 18.

[Rege et al. ()] 'Bipartite isoperimetric graph partitioning for data coclustering'. M Rege , M Dong , F Fotouhi . *Data Min. Knowl. Discov* 2008. 16 (3) p. .

[Dhillon ()] 'Co-clustering documents and words using bipartite spectral graph partitioning'. I S Dhillon . *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, (the 17th ACM SIGKDD international conference on Knowledge discovery and data mining) 2001. p. .

[Shieh et al. ()] 'Collaborative knowledge semantic graph image search'. J.-R Shieh , Y.-T Yeh , C.-H Lin , C.-Y Lin , J.-L Wu . *Proceedings of the 17th international conference on World Wide Web*, (the 17th international conference on World Wide Web) 2008. p. .

[Bauckhage ()] 'Distance-free image retrieval based on stochastic diffusion over bipartite graphs'. C Bauckhage . *Proceedings of the 18th British Machine Vision Conference*, (the 18th British Machine Vision Conference) 2007. p. .

[Sindhwani and Melville ()] 'Document-Word Co-regularization for Semi-supervised Sentiment Analysis'. V Sindhwani , P Melville . *Proceedings of the*, (the) 2008. 2008.

[Eighth IEEE International Conference on Data Mining] *Eighth IEEE International Conference on Data Mining*, p. .

[Schölkopf et al. ()] 'Estimating the Support of a High-Dimensional Distribution'. B Schölkopf , J C Platt , J C Shawe-Taylor , A J Smola , R C Williamson . *Neural Comput* 2001. 13 (7) p. .

[Wang and Cohen ()] 'Languageindependent set expansion of named entities using the web'. R C Wang , W W Cohen . *Proceedings of the 17th IEEE International Conference on Data Mining*, (the 17th IEEE International Conference on Data Mining) 2007. p. .

[Kunegis and Lommatzsch ()] 'Learning spectral graph transformations for link prediction'. J Kunegis , A Lommatzsch . *Proceedings of the 26th Annual International Conference on Machine Learning*, (the 26th Annual International Conference on Machine Learning) 2009. p. .

[Zhou et al. ()] 'Learning with local and global consistency'. D Zhou , O Bousquet , T N Lal , J Weston , B Schölkopf . *Advances in Neural Information Processing Systems*, 2004. 16 p. .

[Manevitz and Yousef ()] 'One-class svms for document classification'. L M Manevitz , M Yousef . *J. Mach. Learn. Res* 2002. 2 p. .

[Yu et al. ()] 'PEBL: Web Page Classification without Negative Examples'. H Yu , J Han , K C Chang , .-C . *IEEE Trans. on Knowl. and Data Eng* 2004. 16 (1) p. .

[Hu et al. ()] 'Preserving Patterns in Bipartite Graph Partitioning'. T Hu , C Qu , C L Tan , S Y Sung , W Zhou . *Proceedings of the 18th IEEE International Conference on Tools with Artificial Intelligence*, (the 18th IEEE International Conference on Tools with Artificial Intelligence) 2006. p. .

[Wang and Hancock ()] 'Probabilistic relaxation labelling using the Fokker-Planck equation'. H Wang , E R Hancock . *Pattern Recogn* 2008. 41 (11) p. .

[Lafferty and Zhai ()] 'Probabilistic relevance models based on document and query generation. In Language Modeling and Information Retrieval'. J Lafferty , C Zhai . *Kluwer International Series on Information Retrieval* 2002.

[Wang et al. ()] 'Probabilistic relevance ranking for collaborative filtering'. J Wang , S Robertson , A P Vries , M J Reinders . *Inf. Retr* 2008. 11 (6) p. .

[Basu et al. ()] 'Recommendation as classification: using social and content-based information in recommendation'. C Basu , H Hirsh , W Cohen . *Proceedings of the 15th national/tenth conference on Artificial intelligence/Innovative applications of artificial intelligence*, (the 15th national/tenth conference on Artificial intelligence/Innovative applications of artificial intelligence) 1998. p. .

[Sun et al. ()] 'Relevance search and anomaly detection in bipartite graphs'. J Sun , H Qu , D Chakrabarti , C Faloutsos . *SIGKDD Explor. Newsl* 2005. 7 (2) p. .

[Yu et al. ()] 'Soft clustering on graphs'. K Yu , S Yu , V Tresp . *Advances in Neural Information Processing Systems 18*, 2005.

[Fern and Brodley ()] 'Solving cluster ensemble problems by bipartite graph partitioning'. X Z Fern , C E Brodley . *Proceedings of the 21st International Conference on Machine learning (ICML04)*, (the 21st International Conference on Machine learning (ICML04)) 2004. p. 36.