Finding Survey Papers via Link and Content Analysis

Ali Daud, Aisha Sikandar and Sameen Mansha

Department of Computer Science and Software Engineering, IIU, Islamabad, 44000, Pakistan
ali.daud@iiu.edu.pk, sikandarayesha@yahoo.com, Sameen_Mansha@yahoo.com

Abstract. Survey articles provide a comprehensive overview of a specific area of research. Automatic detection of survey articles from huge scientific literature is interesting and useful knowledge discovery task in academic social networks. There are different features which can be exploited to differentiate between survey articles and other research articles. Surveys articles are usually citing many important articles this important feature is used in the past for finding surveys using HITS algorithm in addition to base words, base cues, and article length features. The rank of authors writing the articles and text of articles is not considered. In this paper, two additional features based on Author Rank (author authority score of her papers) and textual feature Entropy (paper disorder score) are introduced. Entropy feature has its special significance as it can be used even when there is no link structure. Empirical results show that proposed enhancements are useful and better results are obtained. Especially for large number of top n papers our proposed methods performance is very stable as compared to existing methods.

Keywords: Survey Article Finding, HITS, Author Rank, Entropy, Academic Social Networks

1 Introduction

With the emergence of Web online literature is gathered in many repositories such as DBLP[1] and Citeseer[2] and many academic social network analysis tasks are investigated recently. The co-author and citation based associations between authors and articles, respectively build up these academic social networks. Some interesting tasks are expert finding [12], name disambiguation [17], citation recommendation [13], author interest finding [11] and rising star finding [14]. This work is focused on finding survey articles related to different queries. Survey articles are a type of research articles which provide us detailed literature review about a specific topic in an organized way. They provide researchers a rapid way to jump into new field and save time to first scan and select papers related to a new field. They are useful in grasping the outline of new fields in a short time. Link structure can be exploited to find survey papers.

Previously, automatic detection of survey papers is investigated by [6] using HITS [4] hub scores based on the intuition that survey papers usually cites many important papers in return are candidates of achieving high hub scores. In academic literature survey papers are considered as hubs, while papers initiating new problems, ideas and solutions are considered as authorities, respectively. Important papers are first found and then the papers citing important papers are found as possible candidates of survey papers. Namba and Okumura [6] said that HITS considers links but not content as a result papers with high hub scores even they are not surveys will be detected as survey papers. Consequently, an improved content based HITS algorithm named COMB was proposed [6]. The limitations of COMB is to not work well with sparse link structure and not considering the rank of authors writing papers are raised in this paper. The entropy of paper which is independent of link structure and rank of authors of papers is considered by us in this paper. The intuition is based on the fact that when sparse link structure limits the performance of finding survey papers method the content of paper can be better alternative. It is also important to consider the rank of authors as many survey papers are usually written by the experts in that field with high ranks. Experimental results proved that our proposed methods clearly outperform existing methods for automatic survey finding.

The contributions made in this work are as follows. (1) The usage of paper entropy feature, (2) the usage of the Author Rank score and (3) hybridization of entropy and Author Rank based features for survey paper finding. To the best of our knowledge this is the first work of its nature.

The following paper is organized as follows. Section 2 discusses the related work for HITS, PageRank and Entropy. Section 3 provides the details of features and methods used for finding survey papers. In section 4, dataset, baselines, performance measures, and discussions about the results are given. Section 5 finally concludes this work. The word article and papers is used interchangeably in this paper.

2 Related Work

There is not much work done about survey article finding though HITS is used in only one work. Consequently, we provide literature about the PageRank and Entropy which are used in this work for said task, in addition to HITS. First subsection provides some useful work done by using HITS. In second subsection discussions about PageRank algorithm and its applications are made. Finally in subsection 3 a very useful work of using entropy for ranking authors is discussed.

2.1 HITS Algorithm

HITS algorithm is applied for autonomous citation indexing on the Citeseer corpus[1], a full citation index created by Lawrence et al. [5]. A probabilistic

[1] http://citeseer.ist.psu.edu/

extension named PHITS of HITS algorithm is proposed for ranking paper in Cora corpus[2] with full-text citation index [3]. In both previous work, the full-text papers are automatically categorized into different groups and sorted by their hub or authority scores. A useful application of HITS algorithm is shown for automatic detection of surveys by Namba and Okumura [6]. They said that surveys are the papers which usually cite many important papers so that the papers with high hub scores could be strong candidates of survey papers.

2.2 PageRank Algorithm

PageRank [2] was proposed to rank web pages based on the intuition that the pages linked by many important pages are important. A weakness of PageRank of treating all links equally is raised later and Weighted PageRank algorithm [18] was proposed. It takes both inlinks and outlinks importance into account while calculating the rank of pages and proved effective in retrieving large number of web pages related to a query in comparison to PageRank [6]. Temporal dimension was considered important also in ranking as different events can be popular at different time internals and Time-Weighted PageRank was proposed [16]. It considers page age, event and trend to provide enhanced results.

PageRank was also modified as FolkRank [9] which was used to rank users, tags and resources in Folksonomies on the basis of undirected links between them. Biological networks also benefited from a variation of PageRank named Personalized PageRank which was used to rank proteins by using chemical reactions as directed links between them. Results proved that top ranked proteins are playing important part in human body [10].

2.3 Entropy

Entropy is disorder of a system and was recently used as a counter part of citations to measure the quality of the publication venues (journals, conferences), which is used to rank authors in language models. The lower the entropy the higher the quality of publication venue is considered as the venues with lower entropy usually had high number of citations [15].

3 Survey Article Finding

In this section, the features for survey paper finding are first given with their motivation of usage. Later, the methods used for finding survey papers are explained in detail.

[2] http://www.cs.umass.edu/~mccallum/code-data.html

3.1 Features

This section provides the details of the features used for finding survey papers.

Base Words
It is based on finding specific words in titles of papers for finding survey papers. These words are called base words. They are: survey, review, overview, state-of-the-art and trend.

Base Cues
It is based on finding specific phrases in the content of papers for finding survey papers. These phrases are called base cues. There are two types of phrases used which are positive and negative. Positive phrases are; this survey, this review, this overview, in this survey and we overview and negative phrases are this thesis, this dissertation and we propose.

Size of Paper
Generally, survey papers are longer than others research papers. This means that the papers with more words or sentences have more chances to become survey papers.

HITS
HITS algorithm [4] is a state-of-the-art algorithm used for ranking web pages by calculating authorities and hub scores. It considers two kinds of pages: hubs, which are valuable sources of good links, and authorities, which are valuable because many pages link to them. It is used to rank papers on the basis of papers as nodes and citations providing the directed links graph. The algorithm determines important hub papers in two stages: (1) Constructing directed graph and (2) computing authority and hub score of each paper iteratively (30 iterations in this work) by using following equations:

$$hub(p) = \sum_{q=0}^{n} auth(q)_{q \to p} \quad (1)$$

$$auth(p) = \sum_{q=0}^{n} hub(q)_{p \to q} \quad (2)$$

Where, *hub(p)* is hub score of paper *p* which is the sum of authority scores of all the papers *q* that links to *p*, *auth(p)* is authority score of paper *p* which is sum of hub scores of all the papers *q* to which *p* is linked.

PageRank
PageRank [2] is a state-of-the-art algorithm used for ranking web pages by calculating rank scores of web pages. It is a link analysis algorithm that assigns a numerical weighting to each element of a hyperlinked set of documents, with the purpose of measuring its relative importance within the set. The algorithm may be applied to any collection of entities with reciprocal quotations and references. The

numerical weight that it assigns to any given element *A* is referred to as the PageRank of *A* and denoted by *PR* (*A*). PageRank is used by us to rank authors of papers on the basis of authors as nodes and papers citing each other are providing the directed link graph in this paper.
The algorithm determines rank of the authors in two steps: (1) Constructing directed graph by using paper citations and (2) by computing the page rank score of each author iteratively (30 iterations in this work) using the following equation:

$$PR(PA) = \frac{1-d}{N} + \sum_{p_j \in M(p_i)} PR(p_j)/L(p_j) \quad (3)$$

Where, *PR(PA)* is the PageRank score of one paper of author *A*, *d* is damping factor whose value is 0.85, *N* is total number of papers in the data set, $PR(p_j)/L(p_j)$ is the PageRank score of those papers that links to Paper *A* through citation relationships, divided by number of their out links. One the PageRank score of all paper of an author is calculated, that is summed to get the Author Rank.

Entropy
Entropy is considered as measure of disorder of a system in physics. In this paper, entropy is used to measure the disorder of paper. The unique words in all the papers are noted and term frequency (how many times a single word is present in a paper) and paper frequency (how many times a single word is present in all the documents) is found. And finally entropy is calculated by using the following equation.

$$E(Paper) = \sum_{i=1}^{n} p_i \log_2(p_i) \quad (4)$$

Where, p_i= Term Frequency of word/Document frequency of word.

3.2 Methods

This section provides the existing methods HITS [4] and COMB [6] followed by our proposed methods which are Author Rank Survey Paper Finding, Entropy Survey Article Finding and Three in One Survey Article Finding (3 in 1 SAF), which is combination of all features mentioned in section 3.1.

3.2.1 Existing Methods
In this section an introduction to existing methods HITS and COMB is provided.

HITS
HITS algorithm [4] is applied on the papers to get papers with high hub scores against each query. It is based on the intuition that the survey papers cites many important papers and usually have higher hub scores.

COMB

COMB method [6] is proposed based on five features in combination with HITS which are base words, base cues, size of paper, positional deviation of citations and citation types. Here, for COMB only three features are applied (base words, base cues, and size of paper) in combination with HITS to get top ranked papers against each query by taking into account not only the hub scores of hits as well as the content of papers. The remaining two features are not used as they are shown least significant for survey finding task [6]. The features for COMB are used in a same way they are used for 3 in 1 SAF for comparison purpose.

3.2.2 Proposed Methods

In this section an introduction to our proposed methods Author Rank, Entropy and 3 in 1 SAF are provided.

Author Rank Survey Article Finding

PageRank calculates importance of a page based on the important pages linking to it. Pages are nodes and vertexes represent the links. In case of Author Rank the papers written by authors are the nodes and a paper citing other paper provides directed link to it. The rank for each paper of an author is calculated and then rank of all papers is summed up to get a single value for each author. In this work we have used Author Rank in combination with HITS. Author Rank of an author A_i is compared with the average Author Rank of authors set A, then authority scores are multiplied by 0.5, while the hub scores are doubled to get survey papers against each query.

It is usually thought that the survey papers are written by the highly ranked authors in the field. Conversely, from experiments it is found that the average Author Rank of authors of research papers is greater as compared to Author rank of the authors of survey papers.

Entropy Survey Article Finding

Our proposed entropy feature is merged with HITS in this method. It is usually think that the survey articles are about specific area of research, and then the entropy or disorder in survey paper should have to be less as compared to the research papers. Conversely, form experiments it is found that survey articles have average entropy greater as compared to the research papers. The entropy is calculated using the probabilities of words in papers using standard entropy formula given in Eq. 4.

The entropy E_i of each paper is compared with the average entropy of all papers set E, if entropy of a paper is greater than E; the authority scores are multiplied by 0.5, while the hub scores are doubled to get survey articles against each query.

Three in One Survey Article Finding (3 in 1 SAF)

Our proposed method 3 in 1 SAF comprises of five features in combination with HITS which are base words, base cues, size of paper, Author Rank and entropy of

papers. The Author Rank and entropy is used in 3 in 1 SAF in a similar way as it is explained in above part of our proposed methods subsection. Base words, base cues and size of paper are used in this method in the following way.

Base Words

If base words are present in the title of paper double ($w_{hub)}$ the hub scores of papers and multiply (w_{auth}) authority scores by 0.5 in the opposite case.

Base Cues

The hub scores of research papers is doubled (w_{hub}) if they contain positive cue phrases, and authority score is multiplied by 0.5 (w_{auth}) in the opposite case.

Size of Paper

The length L_i (the number of sentences) of each paper is compared with the average length L, then authority scores are multiplied by sig (L_i/L) (w_{auth}), while hub scores are multiplied by sig (L/L_i) (w_{hub}).

Using the 5 features explained above, we improve HITS algorithm for survey paper detection by taking into account the hub scores, content of papers, rank of authors and entropy of papers. The authority and hub scores of each paper are calculated by the following equations.

$$x_p = \prod_{j=1}^{5} w_{authj} \times \sum_{q \text{ such that } q \to p} y_p \quad (5)$$

$$y_p = \prod_{j=1}^{5} w_{hubj} \times \sum_{q \text{ such that } p \to q} x_q \quad (6)$$

Where, w_{authj} and w_{hubj} indicate 5 weights for authorities and hubs, respectively.

4 Experiments

This section provides the details of the Citeseer dataset, existing methods and performance measures used for comparison. Finally, results and discussions are provided for existing and our proposed methods.

4.1 Dataset

The sample of data is crawled from the Citeseer online computer science publications databases [5]. In total, 20000 papers are taken with 1992 full text papers, 95 survey papers and 32000 unique authors. Papers title or whole text, in-links, out-links, authors and authors in-links and out-links are the used data variables.

Statistical n-gram analysis [7] method is applied on titles of all the papers to get frequency of each phrase. We set a value of 2 and 3 for n. Finally 20 most frequent bi-grams and trigram phrases (queries) with meaningful field in computer science and without overlaps are selected. Later, language model [8] is applied on the titles

of all papers to get matching percentage of papers against each query. Papers are classified against queries on the basis of their high matching score with the query.

Table 1: Queries, number of full text papers and survey papers.

Phrases (Queries)	Full Text Papers	Survey Papers
Bayesian Network	32	1
Support Vector Machine	33	3
Independent Component Analysis	115	4
Data Mining	179	8
AdHoc Network	53	5
Markov Model	88	0
Feature Selection	49	1
Neural Network	434	19
Word Sense Disambiguation	200	4
Information Retrieval	60	4
Image Retrieval	88	12
Machine Learning	144	9
Blind Source Separation	69	0
Fading Channel	38	2
Natural Language Processing	38	0
Object Recognition	29	3
Reinforcement Learning	31	2
Sensor Network	215	13
Speech Recognition	97	5
Average	104.84	5

4.2 Baselines

HITS and COMB are taken as baselines to compare the results with our proposed methods Author Rank, Entropy and 3 in 1 SAF.

4.3 Performance Measures

Precision and recall are the most typical evaluation measures in IR community [1]. F-measure, the harmonic mean of precision and recall, has been used to evaluate the overall performance.

$$\text{Precision} = \frac{\text{number of survey papers correctly detected by a system}}{\text{number of survey papers detected by a system}} \tag{7}$$

$$\text{Recall} = \frac{\text{number of survey papers correctly detected by a system}}{\text{number of survey papers that should be detected}} \tag{8}$$

$$\text{F} - \text{Measure} = \frac{2 \text{ x Recall x Precision}}{\text{Recall + Precision}} \tag{9}$$

4.4 Results and Discussions

It is clear from Figure 1 that for top 5 papers, all methods except Author Rank performs equally well. Though the performance of Author Rank method is poorer for top 5 papers but for different number of top n papers its performance is very stable. For top 10 papers precision for COMB is better as compared to other methods but for greater values of n such as 15, 20 and so on the performance of our proposed methods Author Rank, entropy and 3 in 1 SAF is better as compared to baseline methods.

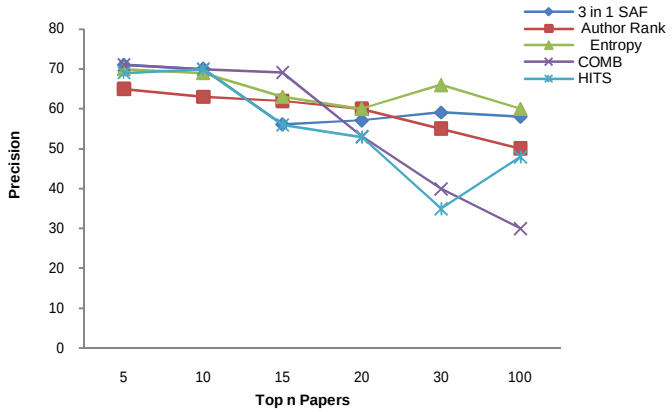

Fig 1: Precision of top n papers

One can see from Figure 2 that for top 5 papers entropy methods performs better as compared to other methods though Author Rank method has also perfumed well. For top 10 papers recall for entropy method is again better though 3 in 1 SAF have comparable performance with it. Just like precision results for greater values of n such as 15, 20 and so on, the performance of our proposed methods Author Rank, entropy and 3 in 1 SAF is better as compared to baseline methods.

From the results shown in Table 2 one can see that by only using HITS the correct results are up to 73.66% for n = 5 and average results are 57.17%. When COMB is used which additionally uses three features the results are improved from 73.66% to 74.33% for n = 5 and average results are 58.81%. But by entropy method the correct results are 75.95% for n = 5 and average results are 69.98, which are much better as compared to the HITS and COMB results. 66% average accurate results for top ranked papers are found by using Author Rank method are also better as compared to simple HITS and COMB results.

For n=5, n=10 and n=15 all the methods perform well, by increasing the value of n, performance of both HITS and COMB become poor i.e. less than 60%. On the other hand Author Rank, Entropy and 3 in 1 SAF performance is still well i.e. greater than 60%.

One can say that a user finding surveys related to query will be interested only in top 5 or 10 papers. Even for this situation our proposed entropy method results are

better as compared to all other methods.

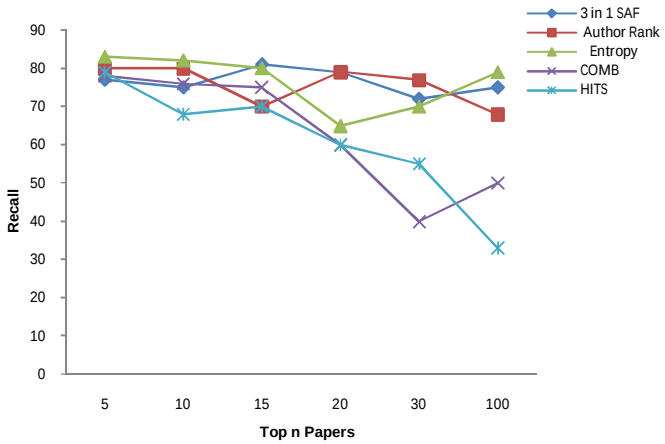

Fig 2: Recall of top n papers

TABLE 2: Evaluation by f-measure of top n papers

Top n papers	Our Proposed Methods			Baselines	
	3 in 1 SAF	Author Rank	Entropy	COMB	HITS
5	73.88	71.73	75.95	74.33	73.66
10	72.42	70.49	74.94	72.87	68.98
15	66.22	65.76	70.49	71.87	62.22
20	66.22	68.21	62.40	56.28	56.28
30	64.85	64.16	67.94	40.00	42.77
100	65.42	57.63	68.21	37.50	39.11
Averages	68.17	66.33	69.98	58.81	57.17

5 Conclusions

The paper addressed the problem of automatic detection of survey papers using entropy and Author Rank features. We can conclude that entropy feature in combination with HITS produced better results. Author Rank feature also produced better results for larger number of top n survey papers when merged with HITS. The COMB and HITS fails when papers links are sparse, but entropy can still work well even there is no link structure. In future, classifiers and learning to rank algorithms can be used for automatically finding survey papers.

Acknowledgments. The work is supported by Higher Education Commission (HEC), Islamabad, Pakistan.

References

1. Azzopardi, L., Girolami, M., and Risjbergen, K.v.: Investigating the relationship between language model perplexity and IR precision-recall measures. In: Proc. of the 26th ACM SIGIR International Conference on Research and Development in Information Retrieval (2003)
2. Brin, S and Page, L.: The anatomy of a large-scale hypertextual Web search engine. In: Proc. of the 7th International Conference on World Wide Web, pp. 107–117 (1998)
3. Cohn, D, and Chang, H.: Learning to probabilistically identify authoritative documents. In: Proc. of the 17th International Conference on Machine Learning, pp.167–174 (2000)
4. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. In: Proc. of the 9th Annual ACM–SIAM Symposium on Discrete Algorithms, pp. 668–677 (1998)
5. Lawrence, S., Giles, L., and Bollacker, K.: Digital libraries and autonomous citation indexing. IEEE Computer, 32(6), pp. 67–71 (1999)
6. Nanba, H., and Okumura, M.: Automatic detection of survey articles. In: Proc. of the 9th International European Conference on Digital Libraries, pp. 391–401 (2005)
7. n-gram, http://en.wikipedia.org/wiki/N-gram.
8. Zhai, C., and Lafferty, J.: A study of smoothing methods for language models applied to ad hoc information retrieval. In: Proc. of the 24th ACM SIGIR International Conference on Research and Development in Information Retrieval, pp. 334–342 (2001)
9 Hotho, A., J¨aschke,R., Schmitz, C., and Stumme, G.: Information Retrieval in Folksonomies: Search and Ranking. In: Proc. of the 3rd European conference on the semantic web: research and applications (ESWC), pp. 411-426 (2006)
10 Ivn, G., and Grolmusz, V.: When the Web Meets the Cell: Using Personalized PageRank for Analyzing Protein Interaction Networks, Biofinformatics, 27(3):405-407 (2011)
11. Daud, A.: Using Time Topic Modeling for Semantics-Based Dynamic Research Interest Finding. Knowledge-Based Systems (KBS), 26:154–163 (2012)
12. Daud, A., Li, J., Zhou, L., and Muhammad, F.: Temporal Expert Finding through Generalized Time Topic Modeling. Knowledge-Based Systems (KBS), 23(6):615-625 (2010)
13. Daud, A., Shaikh, M. A., and Rajpar, A. H.: Scientific Reference Mining using Semantic Information through Topic Modeling. Mehran University Research Journal of Engineering and Technology, 28(2):253-262 (2009)
14. Daud, A., Abbasi, R., and Muhammad, F.: Finding Rising Stars in Social Networks. In: Proc. of International Conference on Database Systems for Advanced Applications (DASFAA). pp. 13-24 (2013)

15. Daud, A. and Hussain, S.: Publication Venue based Language Modeling for Expert Finding. In: Proc. of International Conference on Future Communication and Computer Technology (ICFCCT 2012), 19-20 May (2012)
16. Manaskasemsak, B., Rungsawang, A., and Yamana, H. Time-weighted web authoritative ranking. Information Retrieval Journal 14(2):133-157 (2011)
17. Shu, L., Long, B., Meng, W.: A Latent Topic Model for Complete Entity Resolution. In: Proc. of the International Conference on Data Engineering (ICDE) (2009)
18. Xing, W., and Ghorbani. A.: Weighted PageRank Algorithm. In: Proc. of 2nd Annual Conference on Communication Networks and Services Research, 305-314 (2004)