Proceeding of IC-ITS 2015 e-ISBN:978-967-0850-07-8
International Conference on Information Technology & Society
8-9 June 2015, Kuala Lumpur, MALAYSIA

A Theoretical Framework of Procedures Contain Random Walk Methodology in Various Field of Data Mining and Distributed Computing

Garima Singh Makhija[a], Anjali R. Mahajan[b]

[a]*Wainganga College of Engineering and Management, Nagpur (Maharashtra),India*
[b]*Priyadarshini Institute of Engineering and Technology Nagpur (Maharashtra), India*

garima11makhija21@gmail.com
armahajan@rediffmail.com

Abstract. The concept of Random Walk begins from Social Science and enters into many fields of today like Computer Science, Financial Analysis, and Economics etc. Theory of Random walk is very flexible to understand but at the same time the ambiguity arises to use the same. This paper reflects the use of Random Walk concept in numerous field of Computer Science; it shows that how differently their concepts use in Sensor Network and in the field of Data mining. This paper mainly focused in Random Walk model in Graph based algorithm and shows how it has been used for improving result of previous algorithms and methods given for that particular technique. Mainly, concepts of Data mining which summarized in this paper are Outlier (Isolation Detection) and Clustering (Similarity Detection) which are more prominent research topic in this field. Also, the concept of Target set selection and mobile data gathering in large scale wireless sensor network. Numerous techniques have been developed to these both concepts but the purpose of this paper is to discuss and evaluate algorithms based on Random Walk for specified above methods and their comparison with previous techniques.

Keywords: Random Walk on Graph, Personalized Page Rank, iWander, OutRank.

INTRODUCTION

As the data become information today which is smartly increasing rapidly, in the world of technology, we need an application work interactively and dynamically. A Hybrid combinatorial concept or methodologies are more powerful in computation to answer of an intelligent query rather than Isolated, Static or Hierarchical Methods. If we talk about application related to data Mining field or Distributed in Wireless Sensor network, the concept of Graph is much needed one. The main aim of this paper is to introduce an approach which works on these entire above field and make the computation more powerful i,e the concept of RANDOM WALK ON GRAPH THEORY.A random walk (RW) is a useful model in understanding stochastic processes across a variety of scientific disciplines. This theory supplies the basic probability theory.

Given a graph and a starting point, we select a neighbor of it at random, and move to this neighbor; then we select a neighbor of this point at random, and move to it etc. The (random) sequence of points selected this way is a *random walk* on the graph. A random walk is a finite Markov chain that is time-reversible. [6]. The term random walk was first introduced by Karl Pearson. We are interested in random walks on graphs, where a walker starts from a source node to a destination node and for each step of this travel, the next node to visit is selected uniformly at random from the neighbor-set of the current node. Random walks have been integrated into centrality measurement of social science. For instance, the random-walk betweenness centrality, a relaxation of the shortest-path betweenness. This measure defines how often a node in a graph is visited by a random walker between all possible node pairs. Noh and Rieger [8] introduce the random-walk closeness centrality metric, which measures how fast a node can receive a random-walk message from other nodes in the network.

For distributed approaches of storing and processing of dataset must be present , as due to large scale dataset and graphs are ubiquitous in today's world : www, online social networks, huge search ad query click logs regularly collected and processed by search engine. Because of the massive scale of these datasets, doing analyses and computations on them is infeasible for individual machines. Therefore, there is a growing need for distributed ways of storing and processing these datasets [5]. If we talk about huge search in textual form, Textual Semantic similarity is well known term, in which there is a method to use structured and unstructured knowledge extracted from the English to compute Semantic Similarity. It gives the relation between two documents to hyperlink for detail description on any particular topic, For example, Natural Language Processor , Machine Learning , Data Mining all give detail knowledge on Artificial Intelligence. For more detail, consider these documents and their action in network correspond to them, some of the simplest paths are shown below, but as their length increase there are ofcourse many more:

ML $\xrightarrow{links}$ NLP; where A $\xrightarrow{links}$ B means the text of A contains Hyperlink of B
ML $\xrightarrow{links}$ Artificial Intelligence $\xrightarrow{links}$ NLP;
NLP $\xrightarrow{links}$ Data Mining $\xrightarrow{links}$ ML;
ML $\xrightarrow{cat}$ Mallet (software project) $\xrightarrow{links}$ NLP; where A $\xrightarrow{cat}$ B means A and B belong
to the same document
ML $\xrightarrow{cont:}$ Algorithm $\xrightarrow{links}$ NLP; where A $\xrightarrow{cont}$ B means they have similar lexical contents
NLP $\xrightarrow{cat}$ Grammar Induction $\xrightarrow{links}$ ML.

For the framework many previous approaches have been used like ESA (Explicit Semantic Analysis) but fails due to heavy computation , as it performed disambiguation for all n-grams and compute relatedness of all sense to the context article [2].

There are some other Data Mining approaches like outliers' detection, Clustering, Visualization and many others. Outliers work for real application such as intrusion – detection , the small cluster of outliers often correspond to the interesting events such as denial- of – service or worm attacks. The traditional approaches like density based algorithm shows high detection rate over distance based algorithm for dataset with varying densities, but they can be less effective when identifying small cluster of outliers, to overcome drawback of such algorithm random walk has been used in which it can effectively capture not only the outlying object scattered uniformly but also small cluster of outliers [4].

Another approach related to data mining is highlighted in paper is Website Boundary detection, The nature of the world wide web (or simply the web) is such that information of almost any type can exist, connected in a multitude of ways, which can provide us with information on many subjects. The web is thus a complex interconnected data structure of great diversity. However, there are intuitive notions associated with the information on the web that are used to add meaning to the underlying structure, that are not explicitly described. The notion of a "website" is one of these. The concept of a website is significant with respect to applications such as automatic website map generation, digital preservation and web spam identification. Thus, although there is no agreed definition associated with the term "website", despite its common usage, we are interested in identifying websites. Hence, the Website Boundary Detection (WBD) problem. In the context of the work presented in this paper the WBD problem is defined as follows: The problem of automatically learning the set of web pages/resources/media that are part of a single website [3]. In this problem also they have used advantage of dynamic context over traditional static context; it formed incrementally clusters to produce WBD.

In the field of Wireless computing also it improves many flaws and drawback of traditional approaches. In the problem of data gathering in mobile network, to find clone attack in wireless sensor network and on Wireless sensor Network it is very productive, in mobile data gathering problem, Most contributions that address the data gathering problem in sensor networks focus on distributed data processing techniques combined with broadcast or gossiping protocols for reliable and power-efficient transmission across the network. Some contributions attempt to maximize the energy lifetime of the whole network, instead of the lifetime of each individual sensor, whereas other focus on exploiting the correlation in the data collected by neighboring sensors. Problems related to sensor nodes operating with low duty cycles are investigated e.g. in. Since all of the previous contributions assume a static sensor network that is dense enough to provide at least one path between any sensor node and the data collection point, it is not surprising that mobility is not taken into consideration.

So in this way Random Walk Concept on Graph Theory improves the concept of evaluation dynamically and incrementally from the field of Data Mining to distributed Computing.

DEFINITIONS

Textual Semantic

It compute textual semantic similarity, this approach overcome **Latent Semantic Analysis (LSA)** and Probabilistic LSA , which are unsupervised method that construct a low – dimensional features representation concept space in which words are no longer supposed to be independent, it's only suitable for similarity between two short text fragments because it needs to compare all word pairs. Another used approach without Random walk concept is **Explicit Semantic Analysis (ESA)**, instead of mapping a text to a node (or a small group of nodes) in taxonomy, maps the text to the entire collection of available concepts, by computing the degree of affinity of each concept to the input text. ESA uses Wikipedia articles as a collection of concepts and maps texts to this collection of concepts by use of terms/documents affinity matrix. Similarity is measured in the new concept space, with the implicit (but questionable) assumption that concepts are orthogonal. However, ESA does not use link structure and other structured knowledge from Wikipedia, although these contain valuable information about relatedness between articles.

To overcome all these approaches, the concepts of Random Walk are used to compute various properties of the paths between concept nodes. By using structured and unstructured knowledge extracted from the English version of Wikipedia, A document network built from Wikipedia, capturing various forms of user contributed knowledge, by considering that every article in Wikipedia correspond to a concept node in graph. Relations between concepts are derived from: hyperlinks between articles, lexical content of articles, templates and infoboxes that are invocated, and membership in a category. To estimate the semantic similarity of two text fragments (single words, phrases, sentences, or entire documents), they are first mapped to the vertices of the network built from Wikipedia, and then the distance between sets of vertices is computed [2].

Website Boundary Detection

It reduces the problem of dynamically generated web data in any irrelevant topic, the web is very complex interconnected data structure with great diversity, by using concept of Random Walk graph traversal technique together portion of web data which are the incrementally clustered , to produce WBD solution using fewer web pages than in the case of static context [3].

Outlier Detection

Outlier means something detached from the main body or system, in Data mining outlier detection is anomaly detection, it is the process of finding data objects with behaviour that are very different from expectation. Examples are Fraud detection, security, Industry damage detection, Intrusion detection etc. it basically tries to capture those exceptional cases which substantially deviate from the majority pattern.

In this paper we are going to reflect outlier detection as a stochastic graph- based algorithm using RW called **OutRank** [4], for detecting outlying object, it used the

concept of Markov Chain Model that to built upon the graph, they improved outlying objects by detecting them in the form of clusters which not has been done by previous approaches [4].

Wireless Methodology with Random Walk concept

It is related with data gathering problem in sensor networks focuses on Distributed data processing techniques combined with broadcast or gossiping protocols for reliable powerful efficient transmission across the network. It also introduced Mobile Data Gathering not only with static sensor node but also as a mobile node which patrols the area and collect the desired data [7].

iWander Protocol

A lightweight and distributed protocol to indentify influential user through fixed length RW , it is a distributed for smartphones that leverages RW to identifying influential mobile users, this techniques has been used in many other area also like targeted immunization of infectious diseases , target set selection for information dissemination and many more [1].

OBSERVATIONS

In this section, we review and compare related work about Random Walk concept in Association Technique as Textual Semantic Similarity.

Association Technique as Textual Semantic Similarity

Any concept of study is a network which is made from Wikipedia documents. It uses Random Walk to compute distance between documents; they have made three algorithms**, Hitting /Commute Time, Personalized Page Rank, Truncated Visiting Probability** [2].

Previous Approaches and their Drawback

TABLE(1)

Sr. No.	Technique	Working	Drawback
1	LSA (Latent Semantic Analysis) and Probabilistic LSA	They are unsupervised methods that construct low dimensional representation or “Concept Space” in which words are no longer independent.	It produces larger vocabulary coverage, which are difficult to understand for user to interpret.

2	ESA (Explicit Semantic Analysis)	It maps the text to the entire collection of available concepts instead of map with only nodes; it does not use Wikipedia which is link structure and other structured knowledge which contain valuable information about relatedness between articles. Another approach proposed by Milne and Witten attempt to enrich documents with links to explanatory Wikipedia articles, thus linking structured knowledge to any unstructured text fragments, using a bag of words representation.	It requires heavy computation, as it performs disambiguation for all nodes and computes relatedness of all sense to the context article.

By using Random Walk Method, which approximates text similarity on different data set, first they examine each link separate and measure the effectiveness of Random Walk; they measured Spearman Rank correlation between results of different walks and human judgments based on exact mapping of each word to its closest Wikipedia article. Results of walks on every link type separately are given in Table II.

TABLE II
Spearman Correlation Between Result Of Random Walks And Human Judgments On Wsim353

Link type	Commute time	Personalized page rank	Visiting probability
Hyperlink graph	.654	.664	.684
Content graph	.495	.595	.573
Category graph	.103	.490	.371
Template call graph	.010	.302	.250

In this experiment by algorithms, personalized page rank (PPR) and hitting time are truncated after 5 steps, and every path with probability less than 10^{-5} is truncated when computing visiting probability (VP). In the experiments, the results of PPR and VP were always higher than those of hitting time. It is interesting to find out how much the results of a random walk on different link types are correlated. In Table II we give the Spearman correlation between the scores obtained with VP on different link types. The correlation shows how much the resulting scores based on different link types differ. They have also examined random walk on some different combination of links by choosing the combination weights experimentally and considering the correlation between link types (shown in Table II) and results on each individual link type, with results given in Table III.

TABLE II
Spearman Correlation Between Wsim353 Results Of Visiting Probability On Different Link Types

Link types	Content	Category	Template
Hyperlinks	.7164	.392	.378
Content	–	.374	.517
Category	–	–	.281

The results show that random walk results are improved by combining links in comparison with random walks on individual link types. For example, when equally combining hyperlinks, content links and category links, results are improved in comparison with each link type individually.

TABLE III
Correlation Between Walk Results On Combinations Of Links And Human Judgment For W353 (Ppr: Personalized Page Rank; Vp: Visiting Probability)

w_1 w_2 w_3 w_4	PPR	VP
0.25 0.25 0.25 0.25	.686	.696
0.4 0.4 0.1 0.1	.688	.703
0.3 0.3 0.3 0.1	**.706**	**.707**
(0.7 0.1 0.1 0.1)2 - (0.2 0.6 0.1 0.1)3	.691	.705
(0.2 0.6 0.1 0.1)2 - (0.7 0.1 0.1 0.1)3	.682	.690
(0.5 0.1 0.4 0)2 - (0.3 0.6 0.1 0)3	**.709**	**.711**

CONCLUSION

In this paper we have given theoretical framework for Random Walk on raph , we have presented here different field where Random Walk concepts improve their result in terms of accuracy and flexibility , here we have concentrated on Data Mining technique that also Association by showing Textual Semantic Similarity based on Knowledge extracted from Wikipedia. Results of random walks on different link structures are different, and but combining them gives better results on each task. The similarity measures derived from them can be different as they measure different properties of paths between concepts. Other concepts like Outlier Detection as a Cluster, Website Boundary Detection and Random Walk on Wireless approaches, showing their comparative study is challenging issue that should be investigated more in future by using good and appropriate examples.

ACKNOWLEDGEMENT

We would like to thank to our anonymous reviewers for their constructive and insightful comments.

REFERENCES

1. Bo Han & Aravind Srinivasan. (2012). Your Friend Have More Friends Than You Do: Indentifying Influential Mobile User Through Random Walks, *Mobi Hoc*. 12.
2. Majid Yazdani & Anderi Popescu- Belis. (2010). A Random Walk Framework to Compute Textual Semantic Similarity : a Unified Model for Three Benchmark Tasks, *Semantic Computing (ICSC),424-419.*
3. Ayesh Alshukri & Frans Coenen. (2014). A Dynamic Approach to the Website Boundary Detection Problem Using Random Walks. *Web Intelligence (IAT).*
4. H.D.K. Moonesinghe & Pang-Ning Tan (2006), Outliers Detection Using Random Walks. *Tools with AI.*
5. L. Lovasz. (1993). A Random Walk on Graph: A Survey, *Pual Erdos is Eighty(volume 2).*
6. Luisa Lima & Joao Barros (2007), Random Walk on Sensor Networks,*Symposium on Modelling and optimization in mobile adhoc and wireless networks and workshop.*
7. J. D. Non & H. Rieger. (2004), Random Walk on Complex Network, *Physical Review Letters.*