INTERNATIONAL JOURNAL OF PURE AND APPLIED RESEARCH IN ENGINEERING AND TECHNOLOGY

A PATH FOR HORIZING YOUR INNOVATIVE WORK

SPECIAL ISSUE FOR
INTERNATIONAL LEVEL CONFERENCE
"ADVANCES IN SCIENCE, TECHNOLOGY & MANAGEMENT"
(IC-ASTM)

AN APPLICATION OF GRAPH THEORY IN INFORMATION RETRIEVAL

MRS. DEWASHRI V. MANE

Department of Computer Science, R. L. T. College of Science, Akola, Maharashtra

Accepted Date: 05/09/2017; Published Date: 10/10/2017

Abstract: Graph theory is an important branch of mathematics used to represent relations and network. Graph among all the representation form is a way to present data or knowledge. A graph theory plays a vital role in heterogenic field, one of them is computer science. As the internet is one of the main and best source of information for millions of people, retrieving the information about a particular topic the result will be thousands of web pages related to that topic. The concept of graph theory is used to design web pages, like the link structure of a website could be represented by a directed graph. But our main concern is to discuss the role or the use of graph theory in information retrieval.

Keywords: Graph theory, Konigsberg bridges problem**,** Information retrieval

Corresponding Author: MRS. DEWASHRI V. MANE

Co Author: -

Access Online On:

www.ijpret.com

PAPER-QR CODE

How to Cite This Article:

Dewashri V. Mane, IJPRET, 2017; Volume 6 (2): 117-123

INTRODUCTION

Graph theory is used in almost all the field. The graph theory plays an important role in a wide variety of disciplines, ranging from communications and power systems engineering to molecular and population biology. Graph theory is also widely used in sociology as a way graphs are the most useful mathematical objects in computer science. Especially in research areas of computer science such data mining, image segmentation, clustering, image capturing, networking etc. Graph is a very convenient and natural way of representing the relationship between object, object represent by vertices and the relationship between them by lines. Like link structure of a website could be represented by a directed graph, the vertices are the web pages available at the website and a directed edge from one page to another page exists if and only if there is a link between pages. In same way the most important utilization of graph theory is in information retrieval, such as document retrieval, document classification, collaborative filtering, unified link analysis, image retrieval etc. This paper has been divided into two parts. First part gives the historical background of graph theory and information retrieval. Second part emphasizes how graph theory is used for information retrieval.

Part I: History of Graph Theory

An introduction to basic graph theory through the problem of six people at a party. Robin Wilson, Gresham Professor of Geometry, shows how a good diagram can easily prove that at any party of six people; at least three will either be friends or enemies. In the process, Professor Wilson demonstrates the basics of graph theory. This is the 18th part of 'A Millennium of Mathematical Puzzles'.

The origin of graph theory can be traced back to Euler's work on the Konigsberg bridges problem (1735), which subsequently led to the concept of an Eulerian graph. The study of cycles on polyhedra by the Thomas P. Kirkman (1806 - 95), and William R. Hamilton (1805-65) led to the concept of a Hamiltonian graph.

The concept of a tree, a connected graph without cycles, appeared implicitly in the work of Gustav Kirchhoff (1824-87), who employed graph-theoretical ideas in the calculation of currents in electrical networks or circuits. Later, Arthur Cayley (1821-95), James J. Sylvester(1806-97), George Polya(1887-1985), and others use 'tree' to enumerate chemical molecules.

The study of planar graphs originated in two recreational problems involving the complete graph K_5 and the complete bipartite graph $K_{3,3}$. These graphs proved to be planarity, as was subsequently demonstrated by Kuratowski.

The Konigsberg bridges problem

The Konigsberg bridges problem is the perhaps the best-know example in graph theory. Konigsberg was a city in Russia situated on the Pregel River, which served as the residence of the dukes of Prussia in the 16th century. Today, the city is named Kaliningrad, and is a major industrial and commercial centre of western Russia. The river Pregel flowed through the town, dividing it into four regions, as in the following picture.

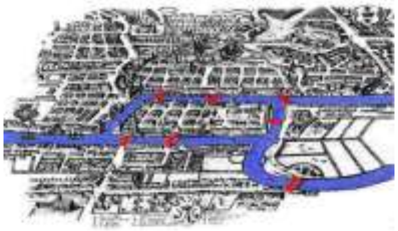

Fig 1 The Konigsberg bridges problem

In the eighteenth century, seven bridges connected the four regions. Konigsberg people used to take long walks through town on Sundays. They wondered whether it was possible to start at one location in the town, travel across all the bridges without crossing any bridge twice and return to the starting point. This problem was first solved by the prolific Swiss mathematician Leonhard Euler, who, as a consequence of his solution invented the branch of mathematics now known as graph theory. Euler's solution consisted of representing the problem by a "graph" with the four regions represented by four vertices and the seven bridges by seven edges as follows:

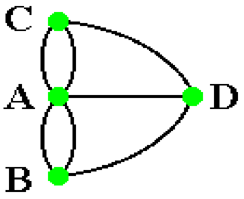

Fig 2 Seven bridges problem

Graph Theory is now a major tool in mathematical research, electrical engineering, computer programming and networking, business administration, sociology, economics, marketing, and communications; the list can go on and on. In particular, many problems can be modeled with paths (see the definition below) formed by traveling along the edges of a certain graph. For instance, problems of efficiently planning routes for mail delivery, garbage pickup, snow removal, diagnostics in computer networks, and others, can be solved using models that involve paths in graphs.

Information Retrieval

Information retrieval (IR) is the science of searching for documents, for information within documents and for metadata about documents, as well as that of searching relational databases and the WWW (World Wide Web). There is overlap in the usage of the terms data retrieval, document retrieval, information retrieval, and text retrieval, but each also has its own body of literature, theory, praxis and technologies. IR is interdisciplinary, based on computer science, mathematics, library science, information science, information architecture, cognitive psychology, linguistics, statistics and physics.

Automated information retrieval system are used to reduce what has been called "information overload". Many universities and public libraries use IR system to provide access to books, journals and other documents. Web search engine are the most visible IR applications. There is vast amount of information to which accurate and speedy access is becoming ever more difficult. One effect of this is that relevant information gets ignored since it is never uncovered, which in turn leads to much duplication of work and effort. With the advent of computers, a great deal of thought has been given to using them to provide rapid and intelligent retrieval systems. In libraries, many of which certainly have an information storage and retrieval problem, some of the more mundane tasks, such as cataloguing and general administration, have successfully been taken over by computers. In principal, information storage and retrieval is simple. Suppose there is a store of documents and a person formulates a question to which the answer is a set of documents satisfying the information need expressed by his question. He can obtain the set by reading all the documents in the store, retaining the relevant documents and discarding all the others. In a sense, this constitutes 'perfect' retrieval.

When high speed computer became available for non-numerical work, many thought that a computer would be able to read an entire document collection to extract the relevant documents. It soon became apparent that using the natural language text of a document not only caused input and storage problem(it still does) but also left unsolved the intellectual problem of characterizing the document content. It is conceivable that future hardware development may make natural language input and storage more feasible. But automatic characterization in which the software attempting to extract information, both syntactic and semantic, from text and using it to decide whether each document is relevant or not to a particular request. The difficulty is to not only knowing how to extract the information but also how to use it to decide relevance. The purpose of an automatic retrieval strategy is to retrieval all the relevant documents at the same time retrieving as few of the non-relevant as possible. When the characterization of a document is worked out, it should be such that when the document it represents is relevant to a query, it will enable the document to be retrieved in response to that query. Human indexers have traditionally characterized documents in this way when assigning index terms to documents.

The indexer attempt to anticipated the kind of index terms a user would employ to retrieval each document whose content is about to describe. Implicitly constructing queries for which the document is relevant. When the indexing is done automatically it is assumed that by pushing the text of a document or query through the same automatic analysis, the output will be a representation of the content. And if the document is relevance to the query, a computational procedure will show this.

Definitions of Information Retrieval on the Web

- ❖ Information retrieval (IR) is the science of searching for documents, for information within documents, and for metadata about documents, as well ... en.wikipedia.org/wiki/Information retrieval
- ❖ The study of systems for indexing, searching, and recalling data, particularly text or other unstructured forms. www.virtechseo.com/seoglossary.htm
- ❖ (IR) part of computer science which studies the retrieval of information (not data) from a collection of written documents. The retrieved documents aim at satisfying a user information need usually expressed in natural language. www.seomoz.org/blog/a-glossary-of-information-retrieval-terminology
- ❖ Searching a body of information for objects that match a search query. www.cs.cornell.edu/wya/DigLib/MS1999/Glossary.html
- ❖ An area of data processing which is concerned with the swift and accurate finding of information in large bodies of data. ... www.uni-due.de/CP/key_terms.htm

Part II : Application of Graph Theory in Information Retrieval

1 Document Retrieval

Traditional Information Retrieval study focuses on modeling the relevance between a textual query and documents. Based on the relevance scoring, a ranked list of documents will be returned in response to the given query. However, the relevance based retrieval does not fully satisfy the information need behind the given query in many occasions.

Two major reasons may attribute to such dissatisfaction:

(i) relevance is hard to be modeled accurately;

(ii) there are factors other than query-document relevance which also govern the satisfaction of information need, such as the importance, topic coverage, etc.

Motivated by improving relevance modeling, in an empirical study is described to testify that hyperlinks overlaps with content similarities to some extent, and an improved TFIDF weighting scheme is proposed in which takes the advantage of the graph structure of web pages. Relevance propagation is also studied to adjust relevance scores by the structure of the document graph. However, much more research effort has been devoted to factors other than relevance that make a better ranking, including algorithms such as PageRank, HITS SALSA , SimRank, implicit link analysis, affinity rank etc.

2 Document Classifications

As a typical application of classification, a number of graph-based semi-supervised learning models have been reported to perform well for text categorization, including Spectral Graph Transducer, harmonic mixture models, learning models on directed graphs. All these studies are “purely graph-based” in that the classification is carried out solely on the graphs, without combining with other textual feature based classification method. A problem with applying graph-based semi-supervised learning models to text categorization is that constructing a graph with a large number of nodes is computationally expensive (which usually involves pairwise similarity computation), while real-world text categorization problems often involves a large number of documents. Besides, some learning algorithms may fail when facing with relatively larger number of examples, such as the manifold learning approach and the harmonic function approach will suffer because the computation of matrix inverse could be prohibitively expensive. However for web page classification, the hyperlink-based graph is more readily available. A few studies have been devoted to improve text-based web page classification by hyperlink graph structures.

3 Collaborative Filtering

Mining user interest patterns from sparse data is a central problem in collaborative filtering. A typical approach is to infer a user pattern from gathering observations on other similar users, and make predictions on an unrated item based on what can be learned from other similar items. Taking this view, the way collaborative filtering works can be explained as letting information propagates along similarity based structures on users and items, which leads to graph-based models. Given the sparseness of the user rating matrix, the problem of developing a reliable similarity measurement between users or items should be solved prior to the prediction of users' interest in unrated items. This problem has been addressed in by introducing different weighting factors and normalization methods to the similarity computation. Recently an algorithm based on maximum margin matrix factorization has been proposed in, which can be also explained from a maximum-margin perspective.

4 Unified Link Analyses

Graph-based models are also referred as link analysis methods in Information Retrieval community. When link analysis becomes popular, various graphs have been constructed for different purposes, which bring a new research topic on unifying multiple graphs. On one hand, multiple graphs can be constructed for a single set of nodes. For example, a Web graph can be built based on different link types, such as hyperlinks, content similarities or user lickthrough. On the other hand, multiple graphs can be constructed for different but closely related sets of nodes. For example, in a retrieval system, three graphs can be build for users, queries and documents respectively. Thus a new problem emerges that how can unify the analysis on those graphs? Several works has been proposed in attempt to unifying the link analysis on multiple graphs. For example, aims at link analysis on multiple graphs for a single set of objects, and in a probabilistic framework are proposed, while aims at a framework to accommodate multiple graphs on interrelated objects.

5 Image Retrieval

It is well-known that low-level features are not adequate for image representation due to its gap with the image semantics. As a result pair wise Euclidean similarity measurement is not reliable, which consequently degrade retrieval performance. Such a problem introduces Manifold earning, which aims at learning a global manifold structure of data from local Euclidean distance, to the image retrieval. For example, uses manifold structure for image representation, and propose manifold ranking schemes. Besides manifold learning approach, there are some other work taking advantage of the graph structure of images, such as where a similarity propagation scheme is proposed over the graph, and where graph-based clustering is performed to facilitate retrieval.

6 Optimal Configurations for Peer-to-Peer User-Private Information Retrieval

User-private information retrieval systems should protect the user's anonymity when performing queries against a database, or they should limit the servers capacity of profiling users. Peer-to-peer user-private information retrieval (P2P UPIR) supplies a practical solution: the users in a group help each other in doing their queries, thereby preserving their privacy without any need of the database to cooperate. One way to implement the P2P UPIR uses combinatoric configurations to administrate the keys needed for the private communication between the peers.

First of all characterize the optimal configurations for the P2P UPIR and see the relationship with the projective planes as described in finite geometry. Then give a very efficient construction of such optimal configurations, i.e.

finite projective planes. Finally check that the involved graphs are Ramanujan graphs, giving an additional justification of the optimality of the constructed configurations.

7 Graph-Based Representations for Similarity Retrieval of Symbolic Images

Image retrieval from an image database by the image objects and their spatial relationships has emerged as an important research subject in these decades. To retrieve images similar to a given query image, retrieval methods must assess the similarity degree between a database image and the query image by the extracted features with acceptable efficiency and effectiveness. This proposes a graph-based model SRG (spatial relation graph) to represent the semantic information of the contained objects and their spatial relationships in an image with no file annotation. In an SRG graph, the image objects are symbolized by the predefined class names as vertices and the spatial relations between object pairs are represented as arcs. The proposed model assesses the similarity degree between two images by calculating the maximum common sub graph of two corresponding SRG's through intersection, which has quadratic time complexity owing to the characteristics of SRG. Its efficiency remains quadratic regardless of the duplication rate of the object symbols. The extended model SRG_T is also proposed, with the same time complexity, for the applications that need to consider the topological relations among objects. A synthetic symbolic image database and an existing image dataset are used in the conducted experiments to verify the performance of the proposed models. The experimental results show that the proposed models have compatible retrieval quality with remarkable efficiency improvements compared with three well-known methods LCS_ Clique, SIM_R, and 2D Be-string, where LCS_ Clique utilizes the number of objects in the maximum common sub image as its similarity function, SIM_R uses accumulation-based similarity function of similar object pairs, and 2D Be-string calculates the similarity of 2D patterns by the linear combination of two 1D similarities.

8 A Graph-Based Object Description for Information Retrieval in Digital Image and Video Libraries

This focuses on the search of a sample object (car) in video sequences and images based on shape similarity. A new description is form for cars, using relational graphs in order to annotate the images where the object of interest (OOI) is present. Query by text can be performed afterward to extract images of OOI from an automatically preprocessed database. The performance of the general retrieval systems is not satisfactory due to the gap between high level concepts and low level features. Successfully fulfill this gap by using the graph-based description scheme which provides an efficient way to obtain high-level semantics from low-level features. Investigate the full potential of the shape matching method based on relational graph of objects with respect to its accuracy, efficiency, and scalability. Here use hierarchical segmentation that increases the accuracy of the detection of the object in the transformed and occluded images. Many shape-based similarity retrieval methods perform well if the initial segmentation is adequate; however, in most cases segmentation without a priori information or user interference yields unsuccessful object extraction results. Compared to other methods, the major advantage of the proposed method is its ability to create semantic segments automatically from the combination of low level edge- or region-based segments using model-based segmentation. It is shown that a graph-based description of the complex objects with model-based segmentation is a powerful scheme for automatic annotation of images and videos.

9 Image Matching Using Alpha-Entropy Measures and Entropic Graphs

Matching a reference image to a secondary image extracted from a database of transformed exemplars constitutes an important image retrieval task. Two related problems are: specification of a general class of discriminatory image features and an appropriate similarity measure to rank the closeness of the query to the database. It present a general method based on matching high dimensional image features, using entropic similarity measures that can be empirically estimated using entropic graphs such as the minimal spanning tree (MST). The entropic measures are generalizations of the well-known Kullback–Liebler (KL) distance, the mutual information (MI) measure, and the Jensen difference. Entropic graph approach has the advantage of being implementable for high dimensional feature spaces for which other entropy-based pattern matching methods are computationally difficult. Now compare this technique to previous entropy matching methods for a variety of continuous and discrete features sets including: single pixel gray levels; tag sub-image features; and independent component analysis (ICA) features. The methodology is illustrating for multimodal face retrieval and ultrasound (US) breast image registration.

Conclusion

The main aim of this paper is a brief overview of the key advances in the field of Information Retrieval, and a description of where the state-of-the-art is at in the field. However, for the task of finding information, these statistical techniques have indeed proven to be the most effective ones so far. Techniques developed in the field have been used in many other areas and have yielded many new technologies which are used by people on an everyday basis, e.g., web search engines, junk-email filters, news clipping services

REFERENCE

1. Shariefuddin Pirzada, Ashay Dharwadker, "Journal of the Korean Society for Industrial and applied Mathematics, Volume 11, No.4,2007
2. Narasingh Deo, "Graph theory with applications to engineering and computer science", Hall of India, 1990.
3. Babai,L., [1977], ' Some applications of graph contractions', J. Graph Theory, Vol. no.1, pp.125-130.
4. Quintana,Y., Kamel, M., Lo,A., [1992]. 'Graph-based retrieval of information in hypertext systems', Proc. of the 10th annual international conference on Systems documentation, pp. 157-168.
5. Kowalski, G., [1983] 'Information Retrieval Systems: Theory and Implementation', Kluwer Stokes, K., Maria Bras-Amorós, [2010] *'Computers & Mathematics with Applications',* Vol. no. 59, pp.1568-1577.
6. Shannon, C.E**., [**1948],' A Mathematical Theory of Communication', Bell System Technical Journal, Vol. no. 27, pp. 379–423, 623–656.
7. Ozer,I. B., Wolf,W., Akansu, A.N., **[**2002], 'Journal of Visual Communication and Image Representation, Vol. no., 13, Pages 425-459.
8. Salton, G., McGill, M., [1983], 'Introduction to Modern Information Retrieval'.