A FRAMEWORK OF GRAPH-BASED TEXTUAL CONTENT REPRESENTATION MODELS IN TEXT MINING

Mr. Vallipagu Reddy Murali [1]
3rd Year Student,
Department of Computer Science,
SV U CM & CS, Tirupati.

Dr. E. Kesavulu Reddy[2],
Assistant Professor,
Department of Computer Science,
SV U CM & CS,, Tirupati.

Abstract: Progress in data innovation has brought a data over stream, with transformative societal ramifications that influence all parts of human life. An extensive and perhaps the most critical bit of this data is as content information, for example, books, news articles, microblogs and texts. These inconceivable amounts of content information must be gotten to and used utilizing PCs, yet the mechanized handling of content is just conceivable utilizing innovation specific for human dialect. Text mining (TM) in an expansive sense alludes to innovation that permits the usage of vast amounts of content information. In the accompanying, this working definition will be revised by a more brief one.

The overview of text mining methodology provides a synthesis of viewpoints on text mining, starting from the linguistic properties and representation of text data, followed by mapping of text mining problems into machine learning tasks, and finally comparing text mining architectures to knowledge discovery processes.

Key Terms : *TM, Data Mining, Graph, Conceptual Graph, n-simple distance.*

I. INTRODUCTION:

The discussion on scalability describes the scalability problem in text mining with examples, implicit views on scalability taken by researchers and practitioners, and existing approaches to scalability.

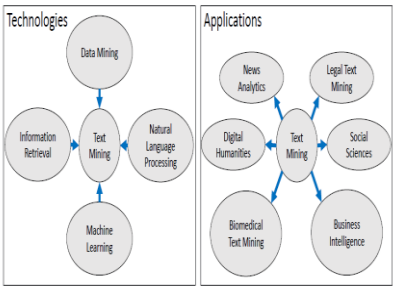

Fig.1: Overview of Technology and Application

In graph textual content representation fashions, a text is represented as a graph containing a fixed of vertices (nodes) and a fixed of edges representing relationships among nodes. even though the use of graphs for representing text has a totally lengthy records in Natural Language Processing (NLP), it has centered on language understanding techniques which include a part of speech tagging, in place of textual content mining duties like textual content classification. Currently, some work considering record class as the goal of graph-based textual content representation techniques has been done. in this bankruptcy, we provide a quick introduction approximately those graph-primarily based models and their application in text class.

Some simple definitions on graphs

A classified graph G is a 4-tuple: $G = (V; E; \alpha; \beta)$, where V is a set of vertices, and $E \subseteq V \times V$ is a fixed of edges that join the vertices, $\alpha: V \rightarrow L_{\upsilon}$, $\beta: V \times V \rightarrow L_e$ are vertices labeling features, and edges labeling features, respectively (with L_{υ} and L_e are the units of labels that can appear on the vertices and edges). We might also seek advice from G as G = (V , E) with the aid of omitting the labeling features.

A graph $G_1 = (V_1; E_1; \alpha_1; \beta_1)$, is a subgraph of a graph $G_2 = (V_2; E_2; \alpha_2; \beta_2)$, denoted $G_1 \subseteq G_2$, if $V_1 \subseteq V_2$, $E_1 \subseteq E_2 \cap (V_1 \times V_1)$, $\alpha_1(x) = \alpha_2(x)$ $_{\upsilon}V_2$ V1, and $\beta_1 (x; y) = \beta_2 (x; y) \forall (x; y) \subseteq E_1$. Conversely, graph G_2 is called a supergraph of G_1.

There are several unique styles of graph. An undirected graph is one graph wherein edges don't have any orientation. Therefore, the brink (a, b) is same to the brink (b, a). In contrast, a graph that has directed edges is known as a directed graph or now and again only a digraph. in the meantime, the time period multigraph refers to a graph wherein a couple of edges between nodes are either accepted or required. another not unusual kind is weighted graph which is a graph wherein each aspect has an related numerical fee, called a weight. typically, the brink weights are non-bad integers. Weighted graphs may be either directed or undirected.

II. COMMON SUB GRAPH MINING

The hassle of FSM can be defined as follows:

"Given a graph dataset D = {G_0, G_1,… G_n} , support(g) denotes the number of graphs (in D) wherein g is a subgraph. The hassle of common subgraph mining is to find any subgraph g such that support(g) ⩾minSup wherein minSup is a minimum aid threshold" [26].

Graph as textual content illustration model

There's a selection of statistics kinds, which may be used to construct graph describing textual content, together with morphological, syntactic, and semantic features. a few fundamental kinds which includes phrase bureaucracy, lemma, stem, a part of speech etc., have carried out commonly in graph fashions. in the meantime, phrase orders, phrase locations or syntax structure are considered as structural statistics. In time period of semantics, several simple semantic information kinds like synonym, hypernym are taken into ac-remember. But, it is quite difficult to capture a deeper semantic meaning of a textual content.

III. GRAPH MODELS FOR INTERNET FILES

Graph models for net documents (or text files in preferred) consisting of 6 methods of creating graphs from internet documents: *general*, *easy*, *n-Distance*, *n-simple Distance*, *Absolute Frequency and Relative Frequency*. All of those graph representations are based on the adjacency of terms in an HTML document.

Standard illustration

The node labels in a document graph are unique because a single node is created for every term, even if a time period appears extra than once in the text. 2D, if a phrase A right now precedes a word B someplace in a section" (textual content material, identify, or link and so forth.) S of the document, then there may be a directed aspect from the node corresponding to time period A to the node corresponding to time period B with an side label B. An area is not created among two phrases if they're separated by way of certain punctuation marks. With this representation, the graph can capture structural information of text (location, relative area of phrases).

There are 3 sections described for fashionable representation consisting of name, hyperlink and text. name consists of the textual content associated with the documents identify and any supplied keywords (metadata). link is the anchor text that looks in links at the record. text contains any of the visible textual content within the file (this includes hyperlinked text, but not the text inside the files name and key phrases). Graph representations are language impartial meaning that they may be applied to a normalized textual content in any language.

An example of a general graph representation for a quick English internet file having the name "SPORT NEWS", a link whose text reads "MORE NEWS", and textual content containing "ENGLAND FOOTBALL NEWS", is proven in Fig. 2, in which TL denotes the name phase, L suggests a hyperlink, and "SPORT","NEWS", "MORE","ENGLAND", "FOOTBALL", which correspond to 5 nodes inside the graph. 4 edges in graph show the family members between words inside the documents: as an example, there's an area from "SPORT" to "NEWS" labeled by means of "TI" meaning that "recreation" at once precedes "NEWS" inside the identify section.

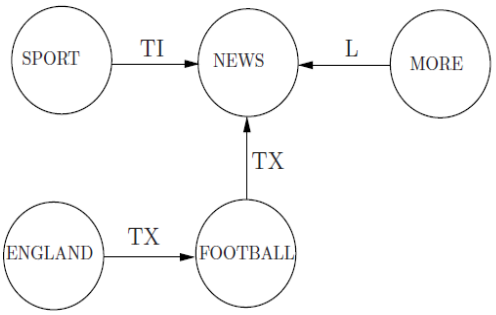

Fig 2: example of a trendy graph illustration of a record

Simple representation

The second one type of Schenker's graph representation is referred to as the simple representation which is basically the same as the same old one, except that no identify or meta-records is tested and the edges in the graph aren't categorized.

n-Distance illustration

The third representation type is n-distance representation. in place of thinking about simplest terms immediately following a given time period in an internet document, we appearance as much as n phrases in advance and connect the succeeding phrases with an aspect this is categorised with the space between them (until the phrases are separated by certain punctuation marks). for instance, inside the graph of the text "ENGLAND FOOTBALL NEWS", there are an edge from "ENGLAND" to football" classified with 1, an part from "ENGLAND" to "NEWS" classified with 2 and an area from "FOOTBALL" to "NEWS" categorised with 1. The graph for this case is shown in figure 2.

n-simple distance representation

The fourth graph illustration, n-easy distance is just like n-distance. this is same to n-distance, but the edges aren't classified that means that we only understand that the gap among related phrases isn't greater than n.

IV. AN ABSOLUTE FREQUENCY ILLUSTRATION

The fifth graph representation is referred to as absolutely the frequency illustration. this is similar to the simple illustration but each node and aspect is labeled with an additional frequency measure. For nodes, this indicates how oftentimes the associated term appeared in the web record. For edges, this shows the variety of times the two linked terms regarded adjacent to each different inside the precise order.

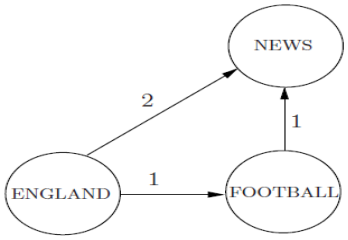

Fig 3: example of a n-distance graph illustration of a file

Relative frequency illustration

The very last graph representation is the relative frequency illustration, that is the same as absolutely the frequency illustration however with normalized frequency values related to the nodes and edges. absolutely the frequency representation uses the total quantity of time period occurrences (at the nodes) and co-occurrences (edges).

An utility in text classification

To calculate distance and similarity measures among graphs for class such as graph edit distance, distance based totally on most common subgraph/minimal not unusual supergraph, state area search technique, probabilistic technique, and many others. By the use of similarity measures among graphs, we will apply several gadget studying techniques (that could work by means of the use of similarity measures between objects) at the graph corpus statistics.

Some complexity

Despite the fact that these graph fashions have the functionality of capturing a few varieties of structural statistics (role, relative place of phrases) in texts, they do now not recall the syntactic structure and semantic family members between phrases.

Hybrid models

So as to overcome the issues of graph models, hybrid models which makes use of common subgraph mining, had been proposed in numerous works. The primary concept of this version is that once representing all documents in Graph-based totally fashions, we use the retrieved graphs to symbolize documents a way that is much like Vector space model: We consider hybrid representation model as a matrix, in which, phrases are columns, files are rows, matrix entries are weights of phrases in documents. Given a corpus of n textual content files C = {d0, d1,.. dn} as input, the stairs to construct a hybrid model are as follows: first off, we represent the text corpus in a graph version like Graph fashions for internet files. After this step, we retrieve a graph corpus G = {G1, G2,… Gn }in which each report di is represented with the aid of a graph Gi.

V. CONCEPTUAL GRAPHS

There are a few rising methods of using greater entire representations of texts than just phrases and simple members of the family between phrases. one of the not unusual strategies to seize the semantic relations between phrases is given by way of Conceptual

In CGs, there are two sorts of nodes which can be standards and members of the family. among them, a Relation node indicates the semantic role of the incident principles. as an instance, the sentence "*Mary is wearing jeans*" can be represented as a conceptual graph as in figure 4. The rectangles and circles in the graph are ideas and members of the family, respectively.

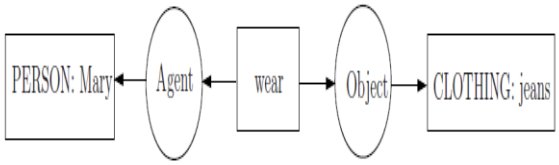

Fig. 4: An instance of Conceptual Graphs

Conceptual Graphs includes wealthy semantic information, so they may be utilized in information representation. A semantic that means of a sentence can be received by means of translating CGs to predicate calculus.

VI. CONCLUSION

1) To begin with, they maintain the important structural information by means of extracting relevant subgraphs from a graph that represents the file.

(2) Secondly, they may be applied in maximum model-primarily based category algorithms for inducing a classification version due to the fact, subsequently; a record is represented via a easy vector. However, the semantic statistics captured in a hybrid version depends at the graph representation used to assemble the hybrid version. within the subsequent phase, we introduce every other form of graph version that has better capability to capture semantic which means.

VII. REFERENCES

[1] Michihiro Kuramochi and George Karypis. Frequent subgraph discovery. In Proceedings of the 2001 IEEE International Conference on Data Mining, ICDM '01, pages 313{320, Washington, DC, USA, 2001. IEEE Computer Society.

[2] A. Markov and M. Last. Model-based classi cation of web documents rep-resented by graphs. In In Proc. of WebKDD 2006: KDD Workshop on Web Mining and Web Usage Analysis, in conjunction with the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006, 2006.

[3] A. Markov, M. Last, and A. Kandel. Fast categorization of web documents represented by graphs. In Proceedings of the 8th Knowledge discovery on the web international conference on Advances in web mining and web usage analysis, WebKDD'06, pages 56{71, Berlin, Heidelberg, 2007. Springer-Verlag.

[4] A. Markov, M. Last, and A. Kandel. The hybrid representation model for web document classi cation. Int. J. Intell. Syst., 23(6):654{679, June 2008}.

[5] Llu s Marquez, Xavier Carreras, Kenneth C. Litkowski, and Suzanne Steven-son. Semantic role labeling: an introduction to the special issue. Comput. Linguist., 34(2):145{159, June 2008.

[6] Sonia Ordo~nez Salinas and Alexander Gelbukh. Information retrieval with a simpli ed conceptual graph-like representation. In Proceedings of the 9th Mexican international conference on Advances in arti cial intelligence: Part I, MICAI'10, pages 92{104, Berlin, Heidelberg, 2010. Springer-Verlag.

[7] Martha Palmer, Daniel Gildea, and Paul Kingsbury. The proposition bank: An annotated corpus of semantic roles. Computational Linguistics, 31(1):71{ 106, March 2005}.

[8] Josef Ruppenhofer, Michael Ellsworth, Miriam R.L. Petruck, Christopher R. Johnson, and Jan Sche czyk. FrameNet II: Extended Theory and Practice. International Computer Science Institute, Berkeley, California, 2006. Dis-tributed with the FrameNet data.

[9] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. Communications of the ACM, 18(11):613{620, 1975}.

[10] A. Schenker, H. Bunke, M. Last, and A. Kandel. Graph-theoretic techniques for Web content mining. Series in Machine P erception and Articial Intelli-gence. World Scientic, 2005.

[11] Michihiro Kuramochi and George Karypis. Frequent subgraph discovery. In Proceedings of the 2001 IEEE International Conference on Data Mining, ICDM '01, pages 313{320, Washington, DC, USA, 2001. IEEE Computer Society.

[12] A. Markov and M. Last. Model-based classi cation of web documents rep-resented by graphs. In In Proc. of WebKDD 2006: KDD Workshop on Web Mining and Web Usage Analysis, in conjunction with the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006, 2006.

[13] A. Markov, M. Last, and A. Kandel. Fast categorization of web documents represented by graphs. In Proceedings of the 8th Knowledge discovery on the web international conference on Advances in web mining and web usage analysis, WebKDD'06, pages 56{71, Berlin, Heidelberg, 2007. Springer-Verlag.

[14] A. Markov, M. Last, and A. Kandel. The hybrid representation model for web document classi cation. Int. J. Intell. Syst., 23(6):654{679, June 2008}.

[15] Llu s Marquez, Xavier Carreras, Kenneth C. Litkowski, and Suzanne Steven-son. Semantic role labeling: an introduction to the special issue. Comput. Linguist., 34(2):145{159, June 2008.

[16] Sonia Ordo~nez Salinas and Alexander Gelbukh. Information retrieval with a simpli ed conceptual graph-like representation. In Proceedings of the 9th Mexican international conference on Advances in artificial intelligence: Part I, MICAI'10, pages 92{104, Berlin, Heidelberg, 2010. Springer-Verlag.

[17] Martha Palmer, Daniel Gildea, and Paul Kingsbury. The proposition bank: An annotated corpus of semantic roles. Computational Linguistics, 31(1):71{ 106, March 2005}.

[18] Josef Ruppenhofer, Michael Ellsworth, Miriam R.L. Petruck, Christopher R. Johnson, and Jan Sche czyk. FrameNet II: Extended Theory and Practice. International Computer Science Institute, Berkeley, California, 2006. Dis-tributed with the FrameNet data.

[19] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. Communications of the ACM, 18(11):613{620, 1975}.

[20] A. Schenker, H. Bunke, M. Last, and A. Kandel. Graph-theoretic techniques for Web content mining. Series in Machine P erception and Articial Intelli-gence. World Scientic, 2005.

[21] Fabrizio Sebastiani. Machine learning in automated text categorization. ACM Comput. Surv., 34(1):1{47, March 2002}.

[22] John F. Sowa. Conceptual Structures: Information Processing in Mind and Machine. Addison-Wesley, 1984.