Graph-based Retrieval Model for Semi-structured Data

Juneyoung Park
Department of Knowledge Service Engineering
Korea Advanced Institute of Science and Technology
Daejeon, Republic of Korea
J.park89@kaist.ac.kr

Mun Y. Yi
Knowledge Service Engineering
Korea Advanced Institute of Science and Technology
Daejeon, Republic of Korea
munyi @kaist.ac.kr

***Abstract*—**The continuous need to process semi-structured data in the more connected and semantic web requires a retrieval model that can truly reflect the user's intention and capture a user's understanding. As a semantic network shows great potential in representing the inherent structure of information in a document, recent studies have attempted to apply semantic networks into information retrieval. While many of the recent works on semi-structured data retrieval focused on the use of field structure within the data. Solely relying on the field structure is insufficient to portray the user's understanding, which is represented through the use of specific query terms. In this study, we seek to overcome this limitation by utilizing a semantic network to model semi-structured data and apply a graph-based semi-structured data retrieval model. Using both a popular testing environment and a real-life query data, we compare the performance of the suggested model with various competitive state-of-the-art retrieval models. The study's findings demonstrate the strength of the proposed model while providing intriguing opportunities for further application of the model.

Keywords— semi-structured retrieval, semantic networks, graph-based retrieval model

I. INTRODUCTION

The rise of connectivity in all things and a more intelligent & semantic web has brought an unavoidable quota of transmission of semantic data, which is often in a semi-structured format [2]. A semi-structured data is formed by various fields with semantic data that naturally describes the property of an object. The information stored within these semantically formed semi-structured data can provide crucial knowledge, thus making the ability to handle such data essential for success.

Semi-structured data, commonly practiced in the XML format, has a steep learning curve and requires much experience in the formal language for fluent use. Evidence provide, the window of opportunity for a prospective approach with a simple and an ad-hoc method to easily access and to navigate structured common data has never been wider. The benefit from such approach would be widely valued in both the professional and independent user community. Through handling these data, a more sophisticated approach to extract the embedded information within the data can be developed and be applied to other tasks such as information retrieval to further enhance the usefulness of semi-structured data.

Graph of text or semantic network can be a suitable approach to capture the rich information of semi-structured data. Graph or network as a representation of reality has been a common and a well-studied area of research[3][10][13]. From web networks to social networks, the benefits of building a network to represent and understand the organic construction of an environment has been widely proven. Semantic network or graph of text is a more specific concept, which represents a document or a text source in a network form. Unlike ontology or rdf graphs, the semantic network is constructed using a naturally occurring text and forms a relationship between keywords in a semantic manner. Recently, there has been a number of researches [3][7][13] that attempt to build and to apply semantic networks to information retrieval. The contextual and comprehensive representation of semantic terms from a text source has provided a valuable contribution to understanding the complex organization of relationships among text.

Specifically, this study investigates a semantic modeling method for semi-structured data using term relationships in two formats: (1) a generic model, which does not require any external resource and (2) a model expanded with additional information from an external resource that has the potential to reflect the genuine relationships between terms in semi-structured data. Our goal is to be able to utilize these two models to provide an enhanced retrieval model for semi-structured data. The contribution of this study is largely in two-folds, first, we identify a generic method that can be utilized for semi-structured data regardless of the domain and without external resource. The generic method will prove that even a generic association between key terms can reflect the intentions of a user's query. Second, we attempt to capture the naturally occurring relationships between terms in an external resource. This is particularly useful as a number of semi-structured data is constructed in regards of multimedia materials available online such as movies, books and other content-based products. The capability of utilizing external resources to enhance the retrieval model will support the goal of capturing the user's intention in a query.

II. Related Works

A. Semi-structured retrieval model

The general approach to semi-structured retrieval models have been through field weight distribution or query reformation. The efforts of early works in field weight distribution such as BM25F[12] and Mixture of Field Language Model(MFLM)[9] utilized a fixed weight for fields across all query terms. The fixed weight was a crucial limitation to the performance of these models as they were unlikely to capture the natural weight of fields by assigning a fixed weight. Recent works by Kim et al[6] introduced the Probabilistic Retrieval Model for Semi-structured data(PRMS) model to address the fixed weight issue by using a term-relevance model based on the probability of a term belonging to a field. Although the intuition behind the PRMS model is comprehensive, there is an inevitable need of a large enough dataset with homogenous field terms that can primarily train the term distribution. Other works focused on reforming a query to better suit semi-structured data. Petkova et al. [11] used the content and the structure of the data in order to transform keyword queries into content-and-structure queries to improve search. Balog et al. [1] explored ways to combine query and category to create a query model.

B. Graph-of-text

A text-based graph, formed with terms or concepts as nodes and their relations as edges, can be formed using relational information ranging from statistical, syntactic, semantic to many more. Specifically, graphs with semantic relations can be described as a thesaurus graph or a concept graph[8][14]. Utilizing these semantic graphs in information retrieval of documents or semantic entity has been a recent interest. The work of Blanco et al[3] and Rousseau et al[13] has been using a graph-specific retrieval model for ad-hoc IR. Similarly, the study by Kim et al[7] transfers the content materials of movie data into a knowledge structure, a form of graph structure, to enhance retrieval effectiveness of movies. Other retrieval models on structured graphs include, a study by Vagena et al[15] suggested a query processing method that builds a twig-query for retrieval on XML tree-graphs. The study by Elbassuoni[5] uses RDF-graphs and sub-graph extraction to retrieve from RDF-graphs using natural keywords.

III. Methodology

In this section, we introduce the methodology to transform a semi-structured data into a semantic network and to build the proximity retrieval model. In all of the following section, we annotate the semantic network G as $G=(V, E)$ where V represents a node in the network and E represents the edge between nodes. A node V is $V=(N, F)$ where N represents the term and F represents the field of which N is included. The edge E is weighed with term proximity between two terms.

A. Generic Semantic Network

The Generic Semantic Network (GSN) is designed to be independent of external resources and formed solely from the semi-structured document. GSN is constructed in two parts, intra-field and inter-field. The intra-field is constructed by the evaluation of proximity between terms within a field. The inter-field is constructed by the evaluation of proximity between terms across fields.

For the intra-field, the proximity score between terms is extracted using a co-occurrence based association evaluation using the concept of Knowledge Structure from Kim et al. [7]. The association between two terms n_1 and n_2 in the same field f can be calculated using the equation shown in the following:

$$Proximity(n_1, n_2) = \frac{\sum co-occur(n_1, n_2)}{Max(Proximity)} \quad (1)$$

The association between the terms n_1 and n_2, $Proximity(n_1,n_2)$, is calculated by the cumulative frequency of co-occurrence of n_1 and n_2 in a sentence within the document. The score is normalized using the maximum proximity score given in the document. The co-occurrence scores of the terms n are collected for each field f.

For inter-field, GSN understands that the separation of fields itself provides a consistent relationship between fields. Therefore, assigns a value alpha to all inter-field associations formally described in the following equation.

$$Proximity(v_{n_1,f_1}, v_{n2,f2}) = \alpha \quad (2)$$

The proximity scores between all nodes in all fields are combined to build a GSN.

B. Wikipedia-based Semantic Network

The Wikipedia-based semantic network (WbSN) differs with the GSN only in inter-field term proximity. WbSN retrieves term proximity inter-field using an external source directly related to the semi-structured data.

The process of term proximity extraction follows a similar process with the intra-field keywords using equation (1). Given the set of keywords in the semi-structured data, the term proximity between keywords is extracted from the Wikipedia page.

C. Proximity Retrieval Model

The proximity retrieval model extracts and cumulates the query term proximity from the respective edges in the semantic network of the target resource. The formal definition of the cumulative score of relevance between a query and the resource is as follows:

$$cumulatedProximity = \frac{\sum_{n1,n2 \subseteq Q \cap E, n1 \neq n2} w_{n1,n2}}{MaxDistance \times Q_n} \quad (3)$$

$$finalScore = EXP(-cumulatedProximity \times \gamma) \quad (4)$$

While n_1 and n_2 are different terms in query Q and document E, $W_{n1,n2}$ represents the edge weight or the shortest distance between terms n_1 and n_2 in the semantic network for document E. MaxDistance represents the longest distance between two terms available in the semantic network of a resource and Q_n represents the length of the query. The cumulated score between terms n_1 and n_2 is normalized using the maximum distance between two terms in a semantic network and the length of the query. The final score in the proximity retrieval

model is smoothed by using an exponential function and a gamma score to capture the linear influence of proximity score.

IV. Experiment & Results

In this section, 2 sets of ad-hoc retrieval tasks are conducted in order to evaluate and demonstrate the performance of GSN in comparison to different semi-structured retrieval models. First, the generalizability of the retrieval model is evaluated through an INEX[1] style evaluation procedure. Second, the performance of GSN and Wikipedia-extended WBSN is evaluated in a more realistic condition, which imitates a real search by a user. Both experiments were based on the INEX-IMDB collection which is publicly available and contains movie documents. The parameters of the various retrieval models were optimized prior to the experiment and the best performing parameter values were used in the experiment.

A. INEX IMDB collection

The INEX-IMDB collection, publicly available for the INEX 2010 & INEX 2011 data-centric track, is a semi-structured representation of the IMDB objects. The collection is formed with many number of fields such as actor, director, producer, genre, release date and country. The INEX-IMDB collection contains 4,418,081 XML documents, which includes 1,594,513 movies and 1,872,471 actors, 129,137 directors, 178,117 producers, and 643,843 others. The test collection was pre-processed(PP) in order to maximize the retrieval performance and understand the real-need of the user, the fields used to index the collection was chosen according to its availableness.

B. Experiment 1

1) INEX Data-Centric Track Topic/Queries

The retrieval models were evaluated using the set of topics and Qrels provided for the INEX Data-Centric track of 2010 and 2011. Total of 63 topics were used in this experiment, a combination of topics from both 2010 & 2011 track. Given a topic T and test collection C, the retrieval models use the topic T to evaluate entries in collection C to retrieve a set of documents $D=d_1,d_2...d_n$ ordered by their relativity to the topic. Each topic is identified by a topic id and formed with a title, castitle, description and narrative. Only the title values were used for the retrieval task as they were the most appropriate to represent a query, which would be used in a real life ad-hoc retrieval task.

2) Evaluations

The effectiveness of all retrieval models were evaluated using two traditional evaluation metrics, Mean Average Precision(MAP) and Precision@1,5,10,30. MAP is a standard and a popular evaluation metrics that produce the mean of average precision for n topics. Precision @ k is defined as the precision score at k returned documents.

3) Retrieval Models

In this experiment, we focus on identifying the general performance of GSN when it is compared to the most popular and well defined model. Using the evaluation metrics described above, the performance of the suggested GSN model is compared with the BM25F model [12]. BM25F is a well-known semi-structured retrieval model and is often the basis of many other variants for semi-structured retrieval tasks.

[1] http://inex.mmci.uni-saarland.de/

Table I Results of Experiment 1

	MAP	P@1	P@5	p@10	p@30
BM25F	0.1281	0.3833	0.3433	0.3017	0.2517
BM25F (PP)	0.1819	0.4667	0.3933	**0.3517†**	**0.2883†**
GSN	**0.2093†**	**0.5†**	**0.4067†**	0.3317	0.2656

4) Results

Results of the retrieval performances are presented in **Table I** while the bold entries represent the best performance in each metric and the † sign represents a statistically meaningful difference (Wilcoxon Test, P <0.01) with the runner-up entry. The results demonstrate that GSN generally outperforms BM25F and is better at higher rank for precision. The general implication of this result portrays the effectiveness of GSN in identifying and allocating the appropriate weight for query keywords and thus provide a higher precision score overall.

C. Experiment 2

1) Amazon Mechanical Turk Queries

A query dataset for the IMDB collection was prepared through crowdsourcing via Amazon Mechanical Turk(MTurk). The use of crowdsourcing in IR has been a recent effort to provide a large and a valuable dataset for retrieval tasks [4]. The participants of MTurk underwent a mock search practice and provided a query consisting multiple keywords to search for a specific movie. In order to guarantee the quality of the queries, there was a strict restriction to only allow users with HIT approval rate greater than 90% in providing the queries. For every query that was accepted, the users were financially rewarded. The final query collection consisted of 6100 queries from 355 users with 3.75 words per query on average.

2) Evaluation Metrics

The performance of the models was evaluated using the metrics MRR and Success@N(s@n). MRR is a standard and popular evaluation metric for retrieval of ranked items, where the value of 1/r is assigned to a query for the rank of target resource r. Success@N assigns a value of 1 when the target resource is recalled within N rank of the retrieved list. Often in a real-life retrieval task, especially for media contents such as movies, a user considers a specific target object and it was deemed appropriate to apply the metrics of Success@n and MRR rather than MAP and Precision measures.

Table II Results of Experiment 2

	MRR	s@1	s@2	s@5	s@10
BM25F	0.64	0.5121	0.64	0.802	0.9011
PRMS	0.64	0.5023	0.6572	0.8323	0.943
CKSM	0.688	0.5807	0.707	0.8438	0.9418
GSN	**0.697†**	0.6354†	0.7446†	**0.8730†**	**0.9449**
WbSN	0.6919††	**0.6946††**	**0.7561††**	0.8053	0.9435

3) Retrieval Models

The retrieval effectiveness of the proposed models GSN and WbSN is compared against several other algorithms including BM25F [12], Probabilistic Retrieval Model for Semi-structured data (PRMS) [6] and Content Knowledge Structure Model (CKSM) [7]. PRMS is a probabilistic retrieval model that guesses the correct field for each query term based on the term distribution across fields and is considered a state-of-the-art algorithm in semi-structured retrieval. The CKSM model uses term proximity in the content of a data, for example the plot of a movie, with a substantially improved performance in content-based retrieval. The CKSM model accumulates the proximity between query terms using only the content of the plot.

4) Results

As shown in **Table II**, the highest performing entry is highlighted in bold. While GSN showed the highest performance for all evaluation metrics against the baseline models. WbSN outperformed all other models, including GSN), at S@1,2. GSN shows on average 3% increase in performance compared to the runner-up model and average 9% increase in performance compared to the least-effective model. The outcome of GSN with statistically significant (Wilcoxon Text, $p<0.01$) increase compared to the runner-up model is indicated with a † while the result entry of WbSN with statistically significant increase (Wilcoxon Text, $p<0.01$) compared to the runner-up (GSN) is indicated with a ††.

The results for WbSN show that there are no significant improvements or differences in the results for higher N. This may be due to the fact that although the term proximity may vary with more accurate estimate of association between terms, the generic term proximity can successfully identify the document to be reasonably relevant. The results with lower N show that the model can more accurately discover the user's intention with a more accurate estimate of the term proximity based on the external resource.

V. CONCLUSION

In this study, we proposed two forms of proximity-based semantic networks, GSN and WbSN, for semi-structured data retrieval. As far as the related works suggest, this study portrays the first attempt to capture semi-structured data into a semantic network format and to apply it to semi-structured data retrieval tasks. Overall, the contribution to utilize a graph-based retrieval model for semi-structured data portrays an intriguing adaptation of extracting human intentions in query via a semantic graph. Naturally, there are a number of encouraging results and opportunities to further develop the retrieval model. The performance evaluations using two different sets of query topics, designed to analyze two different conditions of retrieval effectiveness, show that GSN performs well in both general and specific test conditions. While WbSN, which is built upon GSN but extracts inter-field proximity by utilizing Wikipedia, provides slight improvement for retrieving documents with higher accuracy at lower N, thus suggest the ability of capturing and utilizing a user's intentions in a query.

The evidence not only suggests that capturing the user's intention via term proximity shows great promise but also presents exciting future research opportunities. Given that semi-structured graph data holds a strong potential to specify relevant documents in retrieval tasks using meaningful association between terms, there is a great opportunity to find the most appropriate data resource to accurately represent the real & the relevant association of terms in a semantic network.

ACKNOWLEDGEMENT

This work was supported by Institute for Information & Communications Technology Promotion(IITP) grant funded by the Korea government(MSIP) (No. R2212-15-0027, K-Contents Search/Recommend Service Based on Social Taste Automatic Analysis Platform)

REFERENCES

[1] K. Balog, M. Bron and M. De Rijke. "Query modeling for entity search based on terms, categories, and examples." ACM Transactions on Information Systems (TOIS) 29.4 2011: 22.

[2] C. Bizer, H. Tom, and B. Tim. "Linked data-the story so far." Semantic Services, Interoperability and Web Applications: Emerging Concepts 2009: 205-227.

[3] R. Blanco and L. Christina. "Graph-based term weighting for information retrieval." Information retrieval 15, no. 1 2012: 54-92.

[4] R. Blanco, H. Halpin, DM. Herzig and P. Mika. "Repeatable and reliable search system evaluation using crowdsourcing." Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval. ACM, 2011.

[5] S. Elbassuoni and R. Blanco. "Keyword search over RDF graphs." InProceedings of the 20th ACM international conference on Information and knowledge management, pp. 237-242. ACM, 2011

[6] J. Kim, X. Xue, and W.B. Croft. "A probabilistic retrieval model for semistructured data." Advances in Information Retrieval. Springer Berlin Heidelberg, 2009. 228-239.

[7] S. Kim, K. Han, Mun Y. Yi, S. Cho, and S. Kim. "Exploiting Knowledge Structure for Proximity-aware Movie Retrieval Model." In Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, pp. 1847-1850. ACM, 2014.

[8] AE. Motter, APS. de Moura, YC. Lai and P. Dasgupta. "Topology of the conceptual network of language." Physical Review E 65.6 2002: 065102.

[9] P. Ogilvie and J. Callan. "Combining document representations for known-item search." In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 143-150. ACM, 2003.

[10] L. Page, B. Sergey, M. Rajeev, and W. Terry. "The PageRank citation ranking: bringing order to the Web." 1999.

[11] D. Petkova, W.B. Croft, and Y. Diao. Refining keyword queries for xml retrieval by combining content and structure. In Advances in Information Retrieval, 2009. 662-669. Springer Berlin Heidelberg

[12] S. Robertson, H. Zaragoza and M. Taylor. "Simple BM25 extension to multiple weighted fields." Proceedings of the thirteenth ACM international conference on Information and knowledge management. ACM, 2004.

[13] F. Rousseau and M. Vazirgiannis. "Graph-of-word and TW-IDF: new approach to ad hoc IR." Proceedings of the 22nd ACM international conference on Conference on information & knowledge management. ACM, 2013.

[14] M. Steyvers and J.B. Tenenbaum. "The Large-scale structure of semantic networks: Statistical analyses and a model of semantic growth."Cognitive science 29.1 2005: 41-78.

[15] Z. Vagena, M.M. Moro and V.J. Tsotras. "Twig query processing over graph-structured xml data." In Proceedings of the 7th International Workshop on the Web and Databases: colocated with ACM SIGMOD/PODS 2004, pp. 43-48. ACM, 2004