Entity-centric Summarization: Generating Text Summaries for Graph Snippets

Shruti Chhabra*
«supervised by Srikanta Bedathur»
Indraprastha Institute of Information Technology
New Delhi, India
{shrutic, bedathur}@iiitd.ac.in

ABSTRACT

In recent times, focus of information retrieval community has shifted from traditional keyword-based retrieval to techniques utilizing the semantics in the text. Since such techniques require the understanding of relationships between entities, efforts are ongoing to organize the Web into large entity-relationship graphs. These graphs can be leveraged to answer complex relationship queries. However, most of the research has focused upon extracting structural information between entities such as a path, Steiner tree, or subgraphs. Little attention has been paid to the comprehension of these structural results, which is necessary for the user to understand relationships encapsulated in these structures.

In this doctoral proposal, we pursue the idea of *entity-centric summarization* and propose a novel framework to produce entity-centric summaries which describe the relationships among input entities. We discuss the inherent challenges associated with each module in the framework and present an evaluation plan. Results from our preliminary experiments are encouraging and substantiate the feasibility of summarization problem.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

Keywords

entity-relationship graph, summarization, related entities, relationship description

1. PROBLEM STATEMENT

World Wide Web is a vast source of information and bears the potential to become the largest knowledge base. Significant efforts are being made to add semantics to the web by organizing it into large entity-relationship graphs. Large entity-relationship graphs, such as Dbpedia [4], Freebase [6], and Yago [20], express semantic associations by representing entities as nodes in the graph and relations as edges. Such structures can be exploited to find interesting and complex connections (relations) among entities [24]. Researchers have proposed frameworks to extract relationship structures, ranging from simple paths [16] to complex subgraphs [15], from these entity-relationship graphs. These efforts have considerably improved the semantic search paradigm, however, less attention has been paid to the interpretability of these relationship structures. Due to lack of natural textual descriptions or contextual information in the structures, it becomes difficult to understand and interpret the inherent relationships.

Various modern systems, such as OpenIE [14] and PATTY [31], provide an interface to find relationship between entities and explicitly maintain *textual evidences* from where the relationships were extracted. These evidences are capable of describing the relationships for the entity pair involved. However, even with the availability of evidences, the task of combining the information in evidences for relationships is left to the user.

In this doctoral proposal, we pursue the idea of *entity-centric summarization* i.e., generating textual summaries to describe the relationships among the given set of entities. We assert that presence of entity-centric summaries for entities will help in better understanding of entity relationships. We propose a framework to generate entity-centric summaries for a given set of entities. The set may comprise of one or more entities. In the proposed framework, we broadly address following problems:

- Mining large information sources to extract sentences defining relationships between entities
- Synthesizing a coherent and good quality summary from given set of sentences

1.1 Related Work

Entity-centric summarization is a relatively less explored research problem. Though entity is an essential part of research in information retrieval these days, less emphasis has been laid on generating a summary centered towards a given set of entities. The closest work to the problem proposed is by Srihari *et al.* [35]. They proposed a framework to generate hypothesis graphs (subgraphs connecting given set of entities) and ranked evidence trails (sentences) for the graphs.

[*] research supported by a TCS research fellowship.

WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2567959.

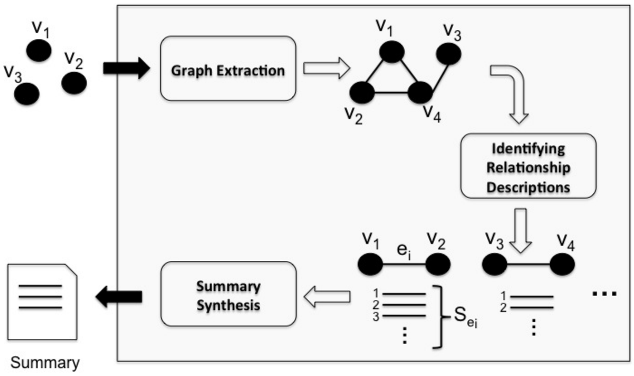

Figure 1: Entity-centric Summarization Framework

Researchers have also looked into some special cases of entity-centric summarization such as generating comprehensive summaries for a domain-specific entity or an entity pair. Sauper *et al.* [34] used high-level structure of human-generated text to automatically create domain specific templates. This research is limited to disease and American film actors entities, which exhibit fairly consistent article structures, thereby providing a good quality template. Filippova *et al.* [17] built a multi-document summarization system to obtain summaries for companies using financial news corpus. Liu *et al.* [29] proposed an integrated bootstrapping model BioSnowball to jointly solve the biography ranking and fact extraction problem. Their system summarizes the web documents to generate Wikipedia-style pages for any person. Jin *et al.* [23] aimed at finding most meaningful evidence trails across documents that connect topics directly or through intermediate topics.

2. PROPOSED FRAMEWORK

Entity-centric summarization refers to generating summaries centered towards a given set of entities. A coherent summary can be generated by identifying and combining textual relationships for the given set of entities. Based on this idea, we propose a novel framework to address the problem of entity-centric summarization. As shown in Figure 1, the framework is composed of three modules namely,

- **Module 1:** Graph Extraction
- **Module 2:** Identifying Relationship Descriptions
- **Module 3:** Summary Synthesis

Firstly, the set of input entities is passed through graph extraction module, where a structural view is generated. The structural view can be seen as a set of edges between connected entities. Next, the sentences describing each edge are extracted from the web corpus, these sentences are termed as *Relationship Descriptions.* In the structural view, each edge is now associated with a set of relationship descriptions. The exhaustive pool of candidate summaries for given entity set is formed by taking cartesian product of relationship description sets. An ordering of these summaries is then obtained using a ranking function. Top-k summaries in the ordering are presented to the user.

2.1 Graph Extraction

Given a set of entities, the task is to extract a subgraph from entity-relationship graph such that the subgraph connects the entities. Let the subgraph be G(V,E), where V represent the set of entities in subgraph and E represents the set of edges i.e., connections among entities. Based on the number of entities in the given entity set, graph extraction module comprises of following two cases:

Case 1: Single entity

Intuitively, in a real life scenario, an entity is generally described by its related entities. For instance, Bt Cotton can be described by entities such as Monsanto - the company which introduced it, Bacillus Thurenginesis - the bacteria which is helpful in creating it, 1996 - the year when it was introduced, and Mexico - the country where it was introduced. Therefore, the goal is to find the salient entities related to the given entity resulting in a star graph like structure. The edges in such a star graph can be augmented with textual evidences, followed by coherently ordering the evidences, to form a potential textual summarization of input entity.

Finding top-k related entities is a well researched problem. TREC[1] 2009, 2010, and 2011 introduced related entity finding (REF) task in its entity track, however, the task provided much more information about the target entity which is generally not available real-time. There are various entity-relationship graphs publicly available, edges in such graphs express semantic associations. Researchers have worked on these graphs to add notion of semantic strength to edges. For instance, Cheng *et al.* [8] proposed a variant of random surfer model to rank the property-value pairs (edges) associated with an entity in a graph. Such techniques may help to identify salient related entities.

Case 2 : Multiple Entities

Given more than one entity, the task is to find a graphical sub-structure connecting these entities. The extracted graph may comprise of additional related entities. For instance, given two entities PDP-1 and Hewlett-Packard, a possible graphical representation may be an entity chain, where PDP-1 is connected to Digital Equipment Corporation (DEC), DEC to Compaq and Compaq to Hewlett-Packard (HP). The relationships among these entities hold

[1] http://trec.nist.gov/data/entity.html

as follows: PDP-1 was introduced by DEC; DEC merged with Compaq, which further merged with HP.

Researchers have addressed the problem of finding relationship structures in various directions. In mid 1980s, Swanson [37] introduced a closed discovery framework for hypothesis generation. Given two disconnected topics, he explored MEDLINE to identify potential linkages via intermediate topics. Srinivasan [36] proposed closed discovery algorithms to automatically return ranked list of intermediate topics, in contrast to manual analysis in original discovery framework. In order to improve effectiveness, researchers [11][21][22] have incorporated domain semantics within the discovery framework. Other than discovery frameworks, researchers have specifically looked into *relationship queries* to identify complex relationships in entity-relationship or entity graphs. Anyanwu and Sheth [3] defined complex relationships such as paths between entities, networks of paths, or subgraphs on RDF as *semantic associations*. Halaschek *et al.* [19] referred the special case of finding path between entities as ρ-path semantic associations and proposed a system SemDIS to discover such semantic associations in RDF. Anyanwu *et al.* [2] further extended the idea to include ρ-join associations in a system, termed as SemRank. Two nodes in RDF are ρ-join associated when they are transitively connected to a common node. Kasneci *et al.* [25] framed the problem of finding relationships between set of two or more entities as a Steiner Tree computation in entity-relationship graphs. Fang *et al.* [16] proposed a system REX to find a subgraph in entity-relationship graph which connects the entity pair. They demonstrated the necessity of including non-paths in the results. Srihari *et al.* [35] proposed a framework to generate ranked list of query relevant hypothesis graphs (subgraphs) from concept-association graph.

As discussed above, significant research efforts have been made to address the problem of sub-graph extraction from a given graph efficiently. Therefore, in this doctoral proposal, we focus on addressing the other two modules of the proposed framework i.e. extracting relationship descriptions and synthesizing a good quality summary.

2.2 Identifying Relationship Descriptions

Each edge $e \in E$ in the graph G(V,E) represents a relationship between entities. We intend to extract set of relationship descriptions (sentences) S_e for each edge e. The task of extracting relationship descriptions appears similar to a sentence retrieval problem. Sentence Retrieval is the task of retrieving a relevant sentence in response to a query, a question, or a reference sentence [30]. However, unlike sentence retrieval problem, our problem deals with a pair of entities as query and aims at retrieving sentences describing relation between them. Some additional research challenges need to be addressed such as defining appropriate ranking measures and considering context.

Various systems such as NELL [7], Open IE [14], and PATTY [31] learn relations from corpus and maintain textual evidences (phrases or sentences) from where relations were extracted. Furthermore, techniques such as support sentence retrieval [5] address the problem of finding sentences for a query (not necessarily an entity) and its associated entities.

Based on the understanding of the problem and literature study, following research challenges are identified:

- Though the retrieved sentence may define a relationship for a given entity pair, the relationship may not be relevant to context of the given graph. Therefore, context must be considered to retrieve relationship descriptions. The context can be defined as adjacent entities in graph or topic of the relationship description.

- An entity may be present in various surface forms which makes the problem of sentence retrieval more complex. For instance, "Mark Zuckerberg" can be mentioned as "Mark E. Zuckerberg", "Mark Elliot Zuckerberg" or "Facebook Creator". Therefore, it is important to consider all mentions of the entity.

2.3 Summary Synthesis

Given description sets $S_{e_1}, S_{e_2}, ..., S_{e_n}$, exhaustive set of candidate summaries is generated by taking Cartesian product of all description sets i.e. $S_{e_1} \times S_{e_2} \times ... \times S_{e_n}$. Thus, a candidate summary is synthesized as follows:

$$l_1 \oplus l_2 \oplus ... \oplus l_n$$

where $l_i \in S_{e_i}$ and $\oplus$ is concatenation operator.

To identify good summaries, the candidate summary set needs to be ranked. The ranking function must capture the inherent properties of a well-written document. Significant research has been made to capture the properties of a well-written document such as Coherence, Focus, Sequentiality, and Non-Redundancy [1][26][33].

Coherence is a vital property of a well-written document which helps the reader to link related pieces of information and comprehend a well connected representation of the text. The text coherence should be considered at two levels: 1) *local coherence*, which implies sentence to sentence transitions should be smooth and 2) *global coherence*, which considers discourse-level relation connecting remote sentences. A number of different theories from a variety of intellectual disciplines have been proposed to represent text coherence in multi-sentence text, including RST, Discourse Grammar, Macrostructures, Coherence Relations, etc [30]. The coherence can be modeled as topical closeness, lexical coherence, temporal coherence, content relatedness, etc.

Sentence ordering also plays a vital role to capture text coherence. The order in which information is presented critically influences the quality of a text and certain orderings may pose problems for the reader trying to understand the gist of the presented information [32]. Focus refers to the property of conciseness, the summary should be *to the point* and should contain only important aspects. A more unified text results in better comprehension [18]. A precise summary should not contain any redundant information [12][13].

Based on the insights from literature, following research issues need to be addressed:

- Ranking function must encapsulate the key properties of a well-written summary. However, it is challenging to create a model considering all properties.

- Ranking all combinations of the relationship descriptions is a computationally expensive task, therefore, pruning step needs to be incorporated to enhance the performance efficiency.

3. EVALUATION PLAN

The generated summaries will be evaluated in following ways:

- The results will be compared with the state-of-the-art technique proposed by Srihari *et al.* [35].
- Top-10 summaries will be graded by human judges on a given scale: Perfect(3), Good(2), Average(1), and Poor(0). Performance scores such as NDCG@k and P@k will be computed for the graded results to measure effectiveness of the ranking algorithm.
- The efficiency of the overall framework will be evaluated using processing time as the measure.
- ROUGE [28], a widely used measure to evaluate summaries, will also be computed against the human generated summaries.
- The ranking function for generated summaries shall be based on various properties of a well written document. Therefore, the summaries will be thoroughly evaluated for each property based on the performance measures inherited from literature.
- A qualitative user study will be conducted to evaluate the usefulness and usability of the entire system.

4. EXPERIMENTS AND RESULTS

Following experiments are performed to identify the underlying challenges in proposed framework:

Experiment 1

The experiment is performed to understand the significance of context sentences in retrieval of relationship descriptions for an entity pair. The experiment is motivated from Blanco *et al.* [5]. They proposed a model for retrieving support sentences for a query and associated entity and their results improved on incorporating context sentences. We also consider the context as two preceding and two succeeding sentences. The input entities e_1 and e_2 are expanded using the *means* relationship in Yago [20]. The sentences are extracted from Wikipedia corpus containing one entity in main sentence and other entity in context or both the entities in the main sentence. The extracted sentences are ranked based on the distance between the two entities. The set of main sentences is referred as Set A. Among the sentences in Set A, sentences containing both the entities in main sentence are filtered (i.e. context ignored) and referred as Set B. Results from Set A and Set B are compared on 127 entity pairs. Precision at rank 5 for Set A and Set B is around 17% and 26% respectively, wheres NDCG for Set A and Set B varies between 90-96%. It clearly indicates that 1) such a definition of context is not helpful in enhancing the performance of sentence retrieval, other formulation of context may be required and 2) better retrieval model is required.

Experiment 2

The purpose of this experiment is to analyze the performance of proposed framework using baseline approaches as well as to determine the key properties of a good quality summary [9]. To demonstrate a basic experiment, we consider the entity chains of path length two (i.e, chain of three entities). We extract the set of relationship descriptions for each edge in graph using Open IE [14]. The experiment is performed on a query set comprising 15 entity chains. The set of candidate summaries are obtained by performing cross product over the identified relationship descriptions for each entity pair. The candidate summaries are ranked based on the cosine similarity distance measure.

The system generated summaries are qualitatively evaluated. The results are encouraging and justify our assertion of proposed framework. High quality summaries generated enabled us to identify three key properties of a good summary: Coherence, Succinctness, and Non-Redundancy.

Experiment 3

The aim of this experiment is to generate a ranked list of summaries for a two-length entity chain [10]. Firstly, each edge in the entity chain is augmented with a set of sentences. The sentences are identified using OpenIE [14]. The cartesian product of the sets of sentences is then computed and each element in the resultant set is considered as a candidate summary. Beside these summaries of length two, it is observed that few single sentences are individually able to describe the three entities in the entity chain. Therefore, single sentences mentioning all the three entities are also considered as candidate summaries. The set of candidate summaries is then ranked based on the properties identified in Experiment 2. The coherence and succinctness properties together are important to form a good quality summary. The less value of any of these may drastically degrade the quality of the summary. Moreover, it is also essential to penalize the summaries with redundant information. Based on this rationale, following function is modeled to rank the summaries:

$$\begin{aligned} Score(m) = {} & (1-\alpha)\ (Coherence * Succinctness) \\ & - \alpha\ \ Redundancy \end{aligned}$$

where m refers to a candidate summary.

In the experiment, coherence is measured using cosine similarity of LSA (latent semantic analysis) [27] vectors. LSA is learned on wikipedia corpus. Succinctness is computed as reciprocal of number of entities in the summary and n-gram overlap is used as a measure to compute Redundancy. The model is evaluated on 27 entity chains. We achieve a high precision of 79% at rank 1 and NDCG of around 95% at rank 5. The results show that summaries generated enable user to understand underlying relationships and also promise for more in-depth work.

5. CONCLUSION

The structural view of relationship among entities are difficult to comprehend due to lack of contextual information, therefore, comprehensive summaries are necessary for better interpretability of results. In this doctoral proposal, we propose *entity-centric summarization* framework to address the challenging problem of generating comprehensive summaries for set of entities. This is one of the few attempts towards comprehending the structural relationships among entities. We briefly discussed the existing approaches and inherent challenges in the proposed framework. The feasibility of generating a textual summary using proposed framework is demonstrated through experiments.

6. REFERENCES

[1] Agrawal, R., Chakraborty, S., Gollapudi, S., Kannan, A., and Kenthapadi, K. Empowering authors to diagnose comprehension burden in textbooks. In *KDD* (2012), pp. 967–975.

[2] Anyanwu, K., Maduko, A., and Sheth, A. Semrank: ranking complex relationship search results on the semantic web. In *WWW* (2005), pp. 117–127.

[3] Anyanwu, K., and Sheth, A. ρ-queries: enabling querying for semantic associations on the semantic web. In *WWW* (2003), pp. 690–699.

[4] Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., and Hellmann, S. Dbpedia - a crystallization point for the web of data. *Web Semant. 7*, 3 (2009), 154–165.

[5] Blanco, R., and Zaragoza, H. Finding support sentences for entities. In *SIGIR* (2010), pp. 339–346.

[6] Bollacker, K., Evans, C., Paritosh, P., Sturge, T., and Taylor, J. Freebase: a collaboratively created graph database for structuring human knowledge. In *SIGMOD* (2008), pp. 1247–1250.

[7] Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Jr., E. R. H., and Mitchell, T. M. Toward an architecture for never-ending language learning. In *AAAI* (2010).

[8] Cheng, G., Tran, T., and Qu, Y. Relin: relatedness and informativeness-based centrality for entity summarization. In *ISWC* (2011), pp. 114–129.

[9] Chhabra, S., and Bedathur, S. Generating text summaries of graph snippets. In *COMAD* (2013), pp. 121–124.

[10] Chhabra, S., and Bedathur, S. Towards generating text summaries for entity chains. In *ECIR* (2014).

[11] Cohen, T., Whitfield, G. K., Schvaneveldt, R. W., Mukund, K., and Rindflesch, T. Epiphanet: an interactive tool to support biomedical discoveries. *JBDC 5* (2010), 21–49.

[12] Dang, V., and Croft, W. B. Diversity by proportionality: an election-based approach to search result diversification. In *SIGIR* (2012), pp. 65–74.

[13] Drosou, M., and Pitoura, E. Search result diversification. *SIGMOD Record 39*, 1 (2010), 41–47.

[14] Etzioni, O., Fader, A., Christensen, J., Soderland, S., and Mausam, M. Open information extraction: the second generation. In *IJCAI* (2011), pp. 3–10.

[15] Faloutsos, C., McCurley, K. S., and Tomkins, A. Fast discovery of connection subgraphs. In *KDD* (2004), pp. 118–127.

[16] Fang, L., Sarma, A. D., Yu, C., and Bohannon, P. Rex: explaining relationships between entity pairs. *VLDB Endowment 5*, 3 (2011).

[17] Filippova, K., Surdeanu, M., Ciaramita, M., and Zaragoza, H. Company-oriented Extractive Summarization of Financial News. In *EACL* (2009), pp. 246–254.

[18] Gray, W. S., and Leary, B. E. What makes a book readable.

[19] Halaschek, C., Aleman-Meza, B., Arpinar, I. B., and Sheth, A. P. Discovering and ranking semantic associations over a large rdf metabase. In *VLDB Endowment* (2004), pp. 1317–1320.

[20] Hoffart, J., Suchanek, F. M., Berberich, K., Lewis-Kelham, E., de Melo, G., and Weikum, G. Yago2: exploring and querying world knowledge in time, space, context, and many languages. In *WWW* (2011), pp. 229–232.

[21] Hristovski, D., Friedman, C., Rindflesch, T. C., and Peterlin, B. Exploiting semantic relations for literature-based discovery. In *AMIA Annu Symp* (2006), vol. 2006, pp. 349–353.

[22] Hristovski, D., Kastrin, A., Peterlin, B., and Rindflesch, T. C. Combining semantic relations and dna microarray data for novel hypotheses generation. In *BioLINK SIG*. 2010, pp. 53–61.

[23] Jin, W., Srihari, R. K., Ho, H. H., and Wu, X. Improving knowledge discovery in document collections through combining text retrieval and link analysis techniques. In *ICDM* (2007), pp. 193–202.

[24] Kasneci, G. Searching and ranking in entity-relationship graphs.

[25] Kasneci, G., Ramanath, M., Sozio, M., Suchanek, F. M., and Weikum, G. Star: Steiner-tree approximation in relationship graphs. In *ICDE* (2009), pp. 868–879.

[26] Kintsch, W., and Van Dijk, T. A. Toward a model of text comprehension and production. *Psychological review 85*, 5 (1978), 363–394.

[27] Landauer, T. K., Foltz, P. W., and Laham, D. An introduction to latent semantic analysis. *Discourse processes 25*, 2-3 (1998), 259–284.

[28] Lin, C.-Y. Rouge: A package for automatic evaluation of summaries. In *ACL* (2004), pp. 74–81.

[29] Liu, X., Nie, Z., Yu, N., and Wen, J.-R. Biosnowball: automated population of wikis. In *KDD* (2010), pp. 969–978.

[30] Mani, I. *Automatic summarization*, vol. 3. John Benjamins Publishing, 2001.

[31] Nakashole, N., Weikum, G., and Suchanek, F. Patty: a taxonomy of relational patterns with semantic types. In *EMNLP-CoNLL* (2012), pp. 1135–1145.

[32] Nenkova, A., and McKeown, K. R. *Automatic summarization*. Now Publishers Inc, 2011.

[33] Pitler, E., and Nenkova, A. Revisiting readability: A unified framework for predicting text quality. In *EMNLP* (2008), pp. 186–195.

[34] Sauper, C., and Barzilay, R. Automatically generating wikipedia articles: a structure-aware approach. In *ACL-IJCNLP* (2009), pp. 208–216.

[35] Srihari, R. K., Xu, L., and Saxena, T. Use of ranked cross document evidence trails for hypothesis generation. In *KDD* (2007), pp. 677–686.

[36] Srinivasan, P. Text mining: generating hypotheses from medline. *JASIST 55*, 5 (2004), 396–413.

[37] Swanson, D. R. Two medical literatures that are logically but not bibliographically connected. *JASIS 38*, 4 (1987), 228–233.