Entity Summarization Based on Formal Concept Analysis

Eun-kyung Kim
School of Computing, KAIST
Daejeon, Republic of Korea
kekeeo@kaist.ac.kr

Key-Sun Choi
School of Computing, KAIST
Daejeon, Republic of Korea
kschoi@kaist.ac.kr

ABSTRACT

This paper presents entity summarization for large-scale knowledge graphs (i.e. a set of *subject-predicate-object* triples) inspired by formal concept analysis. This paper describes extracting tokens from objects and converting a knowledge graph into a formal concept that considers what tokens predicates can take. The extracted concepts naturally form a hierarchical relationship and thus create a graph of objects related to the predicate so that we can determine how the knowledge hierarchy reflects the intrinsic relationships between triples. Our proposed system's effectiveness is illustrated by an experimental study that uses the Entity Summarization Benchmark where we compare our system with six others; this shows that our system outperforms the others with respect to both F-measure and MAP performance measurements.

KEYWORDS

Formal Concept Analysis, DBpedia, Summarization

ACM Reference Format:
Eun-kyung Kim and Key-Sun Choi. 2018. Entity Summarization Based on Formal Concept Analysis. In *Proceedings of International Conference on Information and Knowledge Management (CIKM 2018).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Knowledge Graph (KG) is a large-scale network that describes real-world entities and their relationships. KGs such as DBpedia [5] and LinkedMDB are crucial resources in natural language processing (NLP) applications such as information retrieval, relation extraction, and question answering. KG also plays an important role in implementing artificial intelligence agents for various purposes because it provides access to versatile powerful reasoning techniques to infer and materialize facts as explicit knowledge. Therefore, a huge quantity of facts is constantly being added to the KG.

Typically, the KG is based on a Resource Description Framework (RDF) data model whose knowledge is expressed as$\{(s, p, o)\}$. Here, s, p, and o stand for subject, predicate, and object, respectively. In RDF jargon, subjects and predicates are canonical resources while objects can be either resources or literal values such as strings (e.g. "Mori, Hiroshi") and dates (e.g. "1957-12-07"). Figure 1 shows part of the KG for entity *Hiroshi_Mori_(writer)*.

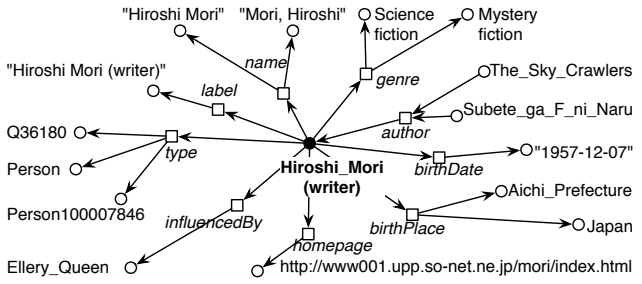

Figure 1: Sample knowledge graph. A subgraph that links all facts related to a specific entity (i.e. Hiroshi_Mori_(writer)) is called entity-graph, and the subjects and objects are represented by circle nodes (the pivot entity "Hiroshi_Mori_(writer)" is a filled circle node) and predicates are square nodes. We can distinguish head (i.e. subject) and tail (object) entities using the directionality of the edges that connect the nodes.

The broad availability of many facts in KG means that people face a disorientation problem with such information. Even worse, structured facts usually cannot distinguish the different meanings of relationships and entities in different triples. Consider the case when you want to find a specific piece of information about an entity you want in KG. An entity in a KG is involved in a set of triple-structured facts that describe it. For example, the latest English version of DBpedia contains 1.7 billion RDF triples for 6.6 million entities. The average of 258 facts per individual entity is not meaningful for quickly identifying an entity's important characteristics and there may be semantically redundant facts. KGs that contain structured facts can be searched using SPARQL, but you will still need to manually scan through each individual triple retrieved by SPARQL until you can identify and understand the entity. This tedious task makes automatic KG summarization very important because users could just read summaries and get an overview of the entity.

Entity summarization is one technique in the area of linked open data for creating a short summary in an entity-graph and has received significant attention in recent years. More specifically, entity summarization is the process of ranking facts in an entity-graph and producing a new short version of the entity-graph without losing any important facts about a given entity. The entity summarization method is designed to help people to quickly identify entities' essential points when searching or browsing a large volume of entity-centric data. Several approaches have been developed to summarize RDF data with respect to entities such as RELIN [1], FACES [3], and LinkSUM [7], but room for further improvement remains.

CIKM 2018, October 2018, Lingotto, Turin, Italy

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

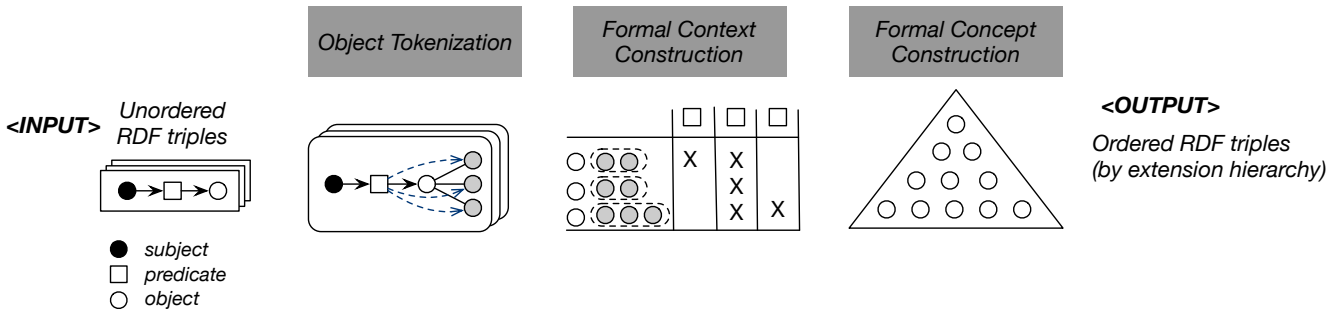

Figure 2: The workflow of proposed system.

This paper proposes yet another approach to summarizing an entity-graph from KGs that is inspired by the Formal Concept Analysis (FCA) approach. FCA was introduced in the early 1980s to formulate concepts and conceptual thinking that could be connected to the philosophical logic of human thought. FCA has been applied in many disciplines such as software engineering, knowledge discovery, data mining, and information retrieval and FCA uses binary object-predicate relationships to construct a knowledge hierarchy that reflects the intrinsic relationships between RDF triples. Our entity summarization system's main goal is to apply FCA to classify RDF triples according to its conceptual importance and rank. The basics of FCA can be found in [2], but we first recall some important definitions and then describe the system in the following section.

2 SYSTEM OVERVIEW

Definition 2.1. (Formal Context $\mathbb{K}$): A formal context is a triple $\mathbb{K} := (G, M, I)$, where G and M are sets and I is the relationship between G and M. The elements of G and M are respectively called **objects** and **attributes**.

This paper refers to "objects" (the elements of G) as "entities" to avoid confusion with *objects* as defined for RDF triples; furthermore, we define a formal concept:

Definition 2.2 (Formal Concept). A formal concept of the formal context $\mathbb{K} := (G, M, I)$ is a pair (A, B) where $A \subseteq G$, $B \subseteq M$, $A' = B$ and $B' = A$. The set A is called the *extent* and B is the *intent* of the formal concept (A, B).

Our proposed system can be described by the workflow presented in Figure 2. The main goal of our entity summarization system is to apply FCA to classify data according to its conceptual importance and rank. The central idea of FCA is the understanding that a fundamental unit of thought is a concept and a context. The proposed system's basic idea is that RDF triples in an entity-graph can be organized in a concept lattice according to the common terms they share. For example, consider the predicate-object pairs set as shown in Table 1. We have 20 different predicate-object pairs in total, and there are a total of 17 entities G (i.e. O_1–O_{17}) and a set of nine attributes M (i.e. P_1–P_9) in the formal context $\mathbb{K} := (G, M, I)$ when duplicates are removed.

The formal context of Table 1 can be represented as a cross-table, as illustrated in Table 2, in which rows are entities and columns are attributes. A formal context is usually given by an incidence matrix where the symbol "X" on line g and column m indicates that object g has attribute m. For example, entity O_{17} ("<http://dbpedia.org/class/yago/JapaneseNovelists>") has attributes P_9 ("<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>"). If the predicate and object are directly connected in pairs in the table, "X" is displayed. Then, we can further separate the entity O_{17} into a series of meaningful tokens (e.g. "Japanese" and "Novelists") containing information to be used to find more relationships between entities and attributes. More specifically, given an entity-graph, the method for constructing the formal context has the following steps:

(1) Set the **objects** of triples as an *entity* of the formal context in rows.
(2) Set the **predicate** to an *attribute* of the formal context in columns.
(3) If a predicate-object pair consists of a single triple, set it to the *relation* of the formal context and mark an "X" in the cell.
(4) If a tokenized unit of entity can be associated with a predicate, set it to the *relation* of a formal context and mark the cell with an "X." For example, O_6, O_7, O_8, O_9, O_{10}, O_{16}, and O_{17}, which contain the token "Japanese" in their object, can share a predicate.

Based on the formal context from Table 2, we can obtain a formal concept like Table 3. FCA aggregates these entities by attributes to form a hierarchy in which $C16$ is located at the bottom, $C1$ is located in the top layer, and $C2$–$C3$ are located in the first layer. As the layer increases, more entities are added while the attributes gradually reduce. The hierarchical structure characterizes the generalization-instantiation relationships between concepts and we can observe that the more entities that are aggregated, the more important the corresponding attribute becomes, but it is also more generalized. Building upon the obtained concepts, we compute the weights of predicate-object pairs using the number of layers in the conceptual hierarchy. After calculating the weights of all predicate-object pairs, we rank them.

3 EXPERIMENTS

This section presents the results of a comparative study between the proposed system and six other baselines (officially posted to the Entity Summarization task). An entity in a knowledge base is involved in a set of triple-structured facts that describe it. When

Table 1: Sample predicate-object pairs from DBpedia entity "Hiroshi_Mori_(writer)"

P_1:=<http://dbpedia.org/ontology/author> O_1:=<http://dbpedia.org/resource/Subete_ga_F_ni_Naru>
P_1:=<http://dbpedia.org/ontology/author> O_2:=<http://dbpedia.org/resource/The_Sky_Crawlers>
P_2:=<http://dbpedia.org/ontology/birthPlace> O_3:=<http://dbpedia.org/resource/Aichi_Prefecture>
P_3:=<http://dbpedia.org/ontology/genre> O_4:=<http://dbpedia.org/resource/Mystery_fiction>
P_3:=<http://dbpedia.org/ontology/genre> O_5:=<http://dbpedia.org/resource/Science_fiction>
P_4:=<http://dbpedia.org/ontology/language> O_6:=<http://dbpedia.org/resource/Japanese_people>
P_5:=<http://dbpedia.org/ontology/nationality> O_6:=<http://dbpedia.org/resource/Japanese_people>
P_6:=<http://dbpedia.org/ontology/notableWork> O_1:=<http://dbpedia.org/resource/Subete_ga_F_ni_Naru>
P_6:=<http://dbpedia.org/ontology/notableWork> O_2:=<http://dbpedia.org/resource/The_Sky_Crawlers>
P_7:=<http://purl.org/dc/elements/1.1/description> O_7:="Japanese writer"@en
P_7:=<http://purl.org/dc/elements/1.1/description> O_8:="Japanese writer"
P_8:=<http://purl.org/dc/terms/subject> O_9:=<http://dbpedia.org/resource/Category:Japanese_mystery_writers>
P_8:=<http://purl.org/dc/terms/subject> O_{10}:=<http://dbpedia.org/resource/Category:Japanese_novelists>
P_8:=<http://purl.org/dc/terms/subject> O_{11}:=<http://dbpedia.org/resource/Category:Living_people>
P_8:=<http://purl.org/dc/terms/subject> O_{12}:=<http://dbpedia.org/resource/Category:Nagoya_University_alumni>
P_8:=<http://purl.org/dc/terms/subject> O_{13}:=<http://dbpedia.org/resource/Category:Nagoya_University_faculty>
P_8:=<http://purl.org/dc/terms/subject> O_{14}:=<http://dbpedia.org/resource/Category:People_from_Aichi_Prefecture>
P_9:=<http://www.w3.org/1999/02/22-rdf-syntax-ns#type> O_{15}:=<http://dbpedia.org/class/yago/Communicator109610660>
P_9:=<http://www.w3.org/1999/02/22-rdf-syntax-ns#type> O_{16}:=<http://dbpedia.org/class/yago/JapaneseMysteryWriters>
P_9:=<http://www.w3.org/1999/02/22-rdf-syntax-ns#type> O_{17}:=<http://dbpedia.org/class/yago/JapaneseNovelists>

Table 2: Formal context example presented as a cross-table representing the relationship between predicates and objects

	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9
O_1	X					X			
O_2	X					X			
O_3	X	X				X		X	
O_4	X		X			X		X	X
O_5	X		X			X			
O_6				X	X		X	X	X
O_7				X	X		X	X	X
O_8				X	X		X	X	X
O_9			X	X	X		X	X	X
O_{10}				X	X		X	X	X
O_{11}				X	X			X	
O_{12}	X					X		X	
O_{13}	X					X		X	
O_{14}	X	X		X	X	X		X	
O_{15}	X					X			X
O_{16}			X	X	X		X	X	X
O_{17}				X	X		X	X	X

there are many facts, entity summarization is the task of selecting a size-constrained subset of triples that best represent the entity. We use the Entity Summarization Benchmark (ESBM) for this shared task; it consists of 140 entities based on two datasets: DBpedia and LinkedMDB. This data set consists of 140 subject entities selected from DBpedia and LinkedMDB. Specifically, there are a total of 100 DBpedia entities consisting of 20 entities of type Agents, Events, Locations, Species, and Works, and 20 entities of entities Films and Persons of LinkedMDB. The benchmark provides each entity's original description to be summarized and its gold-standard summary that was created by crowdsourcing, which are all available as N-Triples documents. To create the gold-standard, six people were added and a summary of lengths 5 and 10 was generated for each given entity's description.

The entity summarization system that uses the method proposed in this paper is called KAIST summarization system using FCA (KAFCA). As a main evaluation metric, we use both F-measure and MAP performance measurements. Tables 4 and 5 present the F-measure and MAP of the seven systems: RELIN, DIVERSUM [6], FACES [3], FACES-E [4], LinkSUM [7], CD [8], and KAFCA, respectively. KAFCA performed better than the other systems with respect to the F-measure performance. In the proposed method, a formal concept was constructed based on the similarity of tokens that constitute value. Therefore, the DBpedia entity summary, which contains a meaningful token in value, performed better. It is difficult to distinguish semantic similarity in the LinkedMDB entity because value is composed of ID in the form of a specific number. The output of the proposed system and all data are available in https://github.com/kekeeo/KAFCA.

4 CONCLUSION

This paper presents KAFCA, a system that generates a summary of entities of a dataset provided by the Entity Summarization Benchmark (ESBM). We assumed that the entity-graph is a collection of concepts, created concepts for each triple, and proposed a way of keeping them connected. This paper considers the problem of detecting influential predicate-objects based on weighted formal concept analysis. We evaluated the efficiency of KAFCA by conducting experiments on Entity Summarization Benchmark and compared KAFCA with several representative influential entity summarization algorithms. These experiments further demonstrated the superiority of KAFCA over state-of-the-art algorithms. In future

Table 3: All formal concepts of Table 2

	Extensions	Intensions
$C1$	$\emptyset$	$\{P_1, P_2, P_3, P_4, P_5, P_6, P_7, P_8, P_9\}$
$C2$	$\{O_{14}\}$	$\{P_1, P_2, P_4, P_5, P_6, P_8\}$
$C3$	$\{O_4\}$	$\{P_1, P_3, P_6, P_8, P_9\}$
$C4$	$\{O_3, O_{14}\}$	$\{P_1, P_2, P_6, P_8\}$
$C5$	$\{O_4, O_5\}$	$\{P_1, P_3, P_6\}$
$C6$	$\{O_4, O_{15}\}$	$\{P_1, P_6, P_9\}$
$C7$	$\{O_9, O_{16}\}$	$\{P_3, P_4, P_5, P_7, P_8, P_9\}$
$C8$	$\{O_4, O_9, O_{16}\}$	$\{P_3, P_8, P_9\}$
$C9$	$\{O_4, O_5, O_9, O_{16}\}$	$\{P_3\}$
$C10$	$\{O_3, O_4, O_{12}, O_{13}, O_{14}\}$	$\{P_1, P_6, P_8\}$
$C11$	$\{O_6, O_7, O_8, O_9, O_{10}, O_{16}, O_{17}\}$	$\{P_4, P_5, P_7, P_8, P_9\}$
$C12$	$\{O_4, O_6, O_7, O_8, O_9, O_{10}, O_{16}, O_{17}\}$	$\{P_8, P_9\}$
$C13$	$\{O_6, O_7, O_8, O_9, O_{10}, O_{11}, O_{14}, O_{16}, O_{17}\}$	$\{P_4, P_5, P_8\}$
$C14$	$\{O_4, O_6, O_7, O_8, O_9, O_{10}, O_{15}, O_{16}, O_{17}\}$	$\{P_9\}$
$C15$	$\{O_1, O_2, O_3, O_4, O_5, O_{12}, O_{13}, O_{14}, O_{15}\}$	$\{P_1, P_6\}$
$C16$	$\{O_3, O_4, O_6, O_7, O_8, O_9, O_{10}, O_{11}, O_{12}, O_{13}, O_{14}, O_{16}, O_{17}\}$	$\{P_8\}$

Table 4: F-measure of selected entity summarizers

Model	DBpedia		LinkedMDB		ALL	
	k=5	k=10	k=5	k=10	k=5	k=10
RELIN	0.250	0.468	0.210	0.260	0.239	0.409
DIVERSUM	0.260	0.522	0.222	0.365	0.249	0.477
FACES	0.272	0.439	0.160	0.259	0.240	0.388
FACES-E	0.285	0.527	**0.252**	0.348	0.276	0.476
LinkSUM	0.290	0.498	0.117	0.255	0.240	0.428
CD	0.299	0.531	0.215	0.326	0.267	0.467
KAFCA	**0.332**	**0.531**	0.249	**0.399**	**0.308**	**0.493**

Table 5: MAP of selected entity summarizers

Model	DBpedia		LinkedMDB		ALL	
	k=5	k=10	k=5	k=10	k=5	k=10
LinkSUM	0.246	0.386	0.120	0.254	0.210	0.348
FACES	0.247	0.386	0.140	0.261	0.216	0.351
DIVERSUM	0.316	0.511	0.269	0.388	0.302	0.476
RELIN	0.348	0.532	0.243	0.337	0.318	0.476
FACES-E	0.354	0.529	0.258	0.361	0.326	0.481
KAFCA	**0.402**	**0.597**	**0.319**	**0.428**	**0.378**	**0.549**

work, we plan to extend KAFCA as a method of calculating the importance of RDF triples with various ranking methods.

ACKNOWLEDGMENT

This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (2013-0-00109, WiseKB: Big data based self-evolving knowledge base and reasoning platform) and KAIST KI Science Technology Leading Primary Research.

REFERENCES

[1] Gong Cheng, Thanh Tran, and Yuzhong Qu. 2011. RELIN: Relatedness and Informativeness-Based Centrality for Entity Summarization.. In *International Semantic Web Conference (1) (Lecture Notes in Computer Science)*, Lora Aroyo, Chris Welty, Harith Alani, Jamie Taylor, Abraham Bernstein, Lalana Kagal, Natasha Fridman Noy, and Eva Blomqvist (Eds.), Vol. 7031. Springer, 114–129. http://dblp.uni-trier.de/db/conf/semweb/iswc2011-1.html#ChengTQ11

[2] B. Ganter and R. Wille. 1999. *Formal Concept Analysis: Mathematical Foundations*. Springer-Verlag.

[3] Kalpa Gunaratna, Krishnaprasad Thirunarayan, and Amit P. Sheth. 2015. FACES: Diversity-Aware Entity Summarization Using Incremental Hierarchical Conceptual Clustering.. In *AAAI*, Blai Bonet and Sven Koenig (Eds.). AAAI Press, 116–122. http://dblp.uni-trier.de/db/conf/aaai/aaai2015.html#GunaratnaTS15

[4] Kalpa Gunaratna, Krishnaprasad Thirunarayan, Amit P. Sheth, and Gong Cheng. 2016. Gleaning Types for Literals in RDF Triples with Application to Entity Summarization.. In *ESWC (Lecture Notes in Computer Science)*, Harald Sack, Eva Blomqvist, Mathieu d'Aquin, Chiara Ghidini, Simone Paolo Ponzetto, and Christoph Lange (Eds.), Vol. 9678. Springer, 85–100. http://dblp.uni-trier.de/db/conf/esws/eswc2016.html#GunaratnaTSC16

[5] Jens Lehmann, Chris Bizer, Georgi Kobilarov, SÃűren Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. 2009. DBpedia - A Crystallization Point for the Web of Data. *Journal of Web Semantics* (2009).

[6] Marcin Sydow, Mariusz Pikula, and Ralf Schenkel. 2013. The notion of diversity in graphical entity summarisation on semantic knowledge graphs. *J. Intell. Inf. Syst.* 41, 2 (2013), 109–149. http://dblp.uni-trier.de/db/journals/jiis/jiis41.html#SydowPS13

[7] Andreas Thalhammer, Nelia Lasierra, and Achim Rettinger. 2016. LinkSUM: Using Link Analysis to Summarize Entity Data.. In *ICWE (Lecture Notes in Computer Science)*, Alessandro Bozzon, Philippe CudrÃĺ-Mauroux, and Cesare Pautasso (Eds.), Vol. 9671. Springer, 244–261. http://dblp.uni-trier.de/db/conf/icwe/icwe2016.html#ThalhammerLR16

[8] Danyun Xu, Liang Zheng, and Yuzhong Qu. 2016. CD at ENSEC 2016: Generating Characteristic and Diverse Entity Summaries.. In *SumPre@ESWC (CEUR Workshop Proceedings)*, Andreas Thalhammer, Gong Cheng, and Kalpa Gunaratna (Eds.), Vol. 1605. CEUR-WS.org. http://dblp.uni-trier.de/db/conf/esws/sumpre2016.html#XuZQ16