The Association Rule Mining System for Acquiring Knowledge of DBpedia from Wikipedia Categories

Jiseong Kim, Eun-Kyung Kim, Yousung Won, Sangha Nam, and Key-Sun Choi

KAIST, Computer Science Department,
Daejeon, Korea
{jiseong,kekeeo,styner0305,nam.sangha,kschoi}@kaist.ac.kr
http://semanticweb.kaist.ac.kr

Abstract. Wikipedia categories are a useful source of knowledge that is usually expressed in a noun-phrase that contains information about concepts of entities or relations among entities. In DBpedia KBs, they categorize their entities into Wikipedia categories using RDF triples. The RDF triples represent only categories of entities, but not concepts of entities or relations among entities despite the fact that expression of Wikipedia categories contain a wealth of those types of information. In this regard, We propose a method that extracts RDF triples encoding concepts of entities or relations among entities from RDF triples encoding Wikipedia categories of each DBpedia entities using association rule mining techniques that mainly utilize lexical patterns in category expression and a hierarchy of categories. Our extensive experiments show that our approach can mine association rules with more high quality than those of state-of-the-art approaches in this problem.

Keywords: Association rule mining, Wikipedia categories, DBpedia enrichment, Knowledge base enrichment

1 Introduction

DBpedia contains plentiful and well-organized knowledge about entities that denote the real world entities like people, animals, and locations or the abstract entities like mathematical concepts and scientific theories. DBpedia uses Wikipedia categories to categorize their entities using RDF triples encoding which DBpedia entities belong to certain Wikipedia categories, which will be called as category triples. For example, the category triple ⟨*dbpedia:John McCarthy, dcterms, category:American computer scientists*⟩ means the DBpedia entity *John McCarthy* belongs to the Wikipedia category *Category:American computer scientists*. Because expression of Wikipedia categories contains information about concepts of entities or relations among entities, we can extract RDF triples encoding those types of information from category triples, which will be called knowledge triples. For example, we can extract the knowledge triples ⟨*dbpedia:John McCarthy, dbpedia-owl:occupation, computer scientist*⟩ from the above mentioned

category triple. This work can be achieved by mining association rules of the form $\{\langle x, belongTo, c\rangle\} \Rightarrow \langle x, r, y\rangle$, which means that if an entity x belongs to the category c, then an entity x and y are in a relation r. We will call these kinds of rules as C2K (category to knowledge) rules. These rules can be mined by the existing association rule mining (ARM) systems like WARMR, ALEPH, and AMIE. But these ARM systems only use frequency information of existing triples in knowledge bases (KBs) to mine rules despite of the fact that there are rich available features like lexical patterns of category expression and dependencies among categories.

In this paper, we propose an effective method to mine C2K rules with fully utilizing lexical patterns in category expression and a hierarchy among categories as features. We compare our method with the state-of-the-art ARM system AMIE that shows outstanding performance on ontological KBs. The experiments show that our method outperforms AMIE in the domain of mining C2K rules. We also propose a simple confidence measure which is more appropriate for mining C2K rules than the standard confidence measure.

In section 2 of this paper, we explore the existing state-of-the art researches that handle the same or similar issues. In the following section 3 and 4, we describe the preliminaries and our approach in much greater details respectively. In section 5, we apply our approach on existing KBs and analyze results in detail. In the last section 6, we conclude and state the future works.

2 Related Work

Knowledge Acquisition from Wikipedia infoboxes. In 2007, Auer *et al.* initiated the DBpedia project [1] that originally extracted knowledge about entities from the Wikipedia infoboxes and encoded it in RDF triples. The project successfully extracted 18M triples from infoboxes, after being further developed. However the fact that only about 44.2% articles have infoboxes results in that only a minor portion of articles are covered by these triples extracted from infoboxes [4]. On the other hand, Wikipedia categories cover about 80.6% articles [4], which means that categories are a rich source of knowledge which is worth being studied in depth.

Knowledge Acquisition from Wikipedia Categories. There have been a number of works done relating to acquire knowledge from Wikipedia categories. For instance, in 2007, Suchanek *et al.* developed YAGO [2, 3], a large ontology derived from Wikipedia categories, infoboxes, and the taxonomic relations of WordNet. YAGO primarily focuses on concepts of entities, i.e., *is-a relation*. They specify the relations (e.g., $locatedIn$) and the corresponding lexical patterns in expression of categories (e.g., Rivers in x) to extract RDF triples encoding concepts of entities or relations among entities. In 2008, Liu *et al.* suggested the approach *Catriple* [4] that analyze lexical patterns in expression of categories using NLP tools and use it to extract knowledge from categories. They enlarged their results by using a subsumption hierarchy contained in Wikipedia category network. In 2008, Nastase and Strude suggested a similar approach

[5] that classifies each category into several classes based on how clear they express conceptual or relational information about entities. They analyze lexical patterns in each class of categories using NLP tools and the category network, and then use it to extract knowledge from categories. In this paper, we mainly focuses on an association rule mining (ARM) system that mines C2K rules more effectively than other state-of-the-art ARM systems, so we do not compare our final prediction to triples of the above-mentioned three approaches.

Association Rule Mining Systems. Association rules are mined on a list of transactions. A transaction is a set of items. For example, in the context of sales analysis, an item is a product and a transaction is a set of products bought together by a customer in a specific event. The mined association rules are of the form $\{Milk, Diaper\} \Rightarrow Beer$, meaning that people who bought a bottle of milk and a diaper usually also bought a beer, which is partly because of the fact that working couples with children are so busy to go to a beer bar. These association rules can be used to discover knowledge about entities in KBs. For example, we can mine rules with the form $\{\langle e_1, r_1, e_2\rangle, \langle e_2, r_2, e_3\rangle\} \Rightarrow \langle e_1, r_3, e_3\rangle$ where e is one of entities, which means that if there are $\langle e_1, r_1, e_2\rangle$ and $\langle e_2, r_2, e_3\rangle$ in KBs, $\langle e_1, r_3, e_3\rangle$ is likely to exist in KBs. We can predict new triples using these rules. In 2013, Galarraga suggested the state-of-the-art ARM system AMIE [6] that mines association rules from KBs with more high scalability and quality of results than the previous ARM systems WARMR [7–9] and ALEPH[1]. AMIE can achieve high performance by using their new confidence measure and the efficient algorithm for mining rules in KBs. Despite of their superior capability, they have weaknesses in the more specific problem, mining C2K rules that have the form $\{\langle x, belongTo, c\rangle\} \Rightarrow \langle x, r, y\rangle$, because they only use frequency information of existing triples in KBs to mine rules. Wikipedia categories are usually expressed in a noun phrase and organized in a network, so there are plentiful lexical and hierarchical information that we can further utilize to solve this problem. Our approach uses these kinds of information to be a specialist in the domain of mining C2K rules.

3 Preliminaries

DBpedia: The RDF Knowledge Base. In this paper, we focus on DBpedia, one of RDF knowledge bases (KBs). An RDF KB is a set of RDF triples that encode relations between an entity and a literal (a string, a integer, a date and so on) or relations among entities. Each RDF triple has the form $\langle s, p, o\rangle$ with s denoting the subject which is placed with one of entities in KBs, p denoting the predicate which represents one of relations pre-defined in KBs, and o the object which is placed with one of entities in KBs or literals.

Wikipedia Categories in DBpedia. DBpedia contains RDF triples encoding which DBpedia entities belong to certain Wikipedia categories. We will call these triples as **category triples**. Each category triple has the form $\langle e, r_{cat}, c\rangle$

[1] `http://www.cs.ox.ac.uk/activities/machlearn/Aleph/aleph_toc.html`

with e denoting one of entities in KBs, r_{cat} denoting the relation indicates that the subject is categorized in the object, and c denoting one of Wikipedia categories.

Types of Categories. There are different types of Wikipedia categories: One is *the conceptual categories* that contain information about a class of an entity (e.g., Albert Einstein is in the category *Category:Naturalized citizens of the United States*). The other is *the relational categories* (like *Category:1989 births*) containing relational information with other entities or literals. There are some other types that merely indicate thematic vicinity (e.g., *Category:Physics*). Of these, *the conceptual categories* and *the relational categories* are main sources of knowledge for our method to extract information about entities.

Category Expression. Wikipedia categories are usually expressed in a noun phrase which is composed of nouns with some modifiers (prepositional phrases, adjectives and so on) and some special characters (e.g., hyphens(-), commas(,) and so on). Figure 1 below shows examples of Wikipedia categories. We can see that there is plentiful information about occupations, locations, educations and so on. Our method analyzes lexical patterns in expression of categories and use it to extract knowledge.

20th-century mathematicians
Philosophers who committed suicide
People of Rhode Island in the American Civil War
Alumni of King's College, Cambridge

Fig. 1. The examples of Wikipedia categories

Category Hierarchy. Since Wikipedia categories are more compressively expressed with a few lexical elements than a usual complete sentence, lexical patterns are not sufficient to get high quality results. In this regard, we use a category hierarchy extracted from the Wikipedia category network to enhance our results. The Wikipedia category network (N_{cat}) is a directed graph which encodes various dependencies (subsumption relationships, semantic similarity, thematic similarity and so on) among categories. Our method mainly use subsumption relationships between categories, which are partly contained in N_{cat}. To reduce errors introduced by dependencies representing non-subsumption relationships among categories, we convert N_{cat} to a directed acyclic graph, which will be called as a category hierarchy (H_{cat}), by eliminating cycles in N_{cat}. Although H_{cat} is not a complete subsumption hierarchy, but it is useful for our method.

Association Rules and C2K Rules. An association rule consists of a body and a head, where the body is a set of atoms and the head is a single atom which is an RDF triple that can have variables at the subject and/or object position. For example, in the association rule $\{\langle x, r_1, y\rangle\} \Rightarrow \langle x, r_2, y\rangle$, the set of the atom $\{\langle x, r_1, y\rangle\}$ is the body of the rule and the atom $\langle x, r_2, y\rangle$ is the head of the rule. In this paper, we only focus on **C2K (category to knowledge) rules** whose

body is composed of one category triple with a subject-position variable and head is composed of one knowledge triple, that encodes concepts of entities or relations among entities, with a subject-position variable. In the next section, we will define our problem more formally.

The C2K Problem. Let E be a set of entities, R denotes a set of relations, L denotes a set of literals and C denotes a set of categories of entities. A set of n category triples can be represented as $T_{cat} = \{\langle e_i, r_{cat}, c_i\rangle\}_{i=1}^{n}$ where $e \in E$, $r_{cat} \in R$ and $c \in C$. A set of m knowledge triples can be represented as $T_{know} = \{\langle e_i, r_i, o_i\rangle\}_{i=1}^{m}$ where $e \in E$, $r \in R - \{r_{cat}\}$ and $o \in E \cup L$. An RDF KB containing n category triples and m knowledge triples can be represented as $K = T_{cat} \cup T_{know}$. The problem is to mine C2K rules of the form $t_{cat}^{x} \Rightarrow t_{fact}^{x}$ from K, where $t_{cat}^{x} = \langle x, r_{cat}, c \in C\rangle$, $t_{fact}^{x} = \langle x, r \in R - \{r_{cat}\}, o \in E \cup L\rangle$, and x is a variable for an entity $e \in E$.

Goal. The goal of our approach is to solve the C2K problem with DBpedia KB and Wikipedia categories to mine a wealth of C2K rules with high quality.

4 The Proposed Approach

For mining C2K rules, our method mainly use lexical patterns in expression of categories and a hierarchy of categories, i.e., H_{cat}, which is extracted from Wikipedia category network. In the first step of our approach, we discover lexical patterns of each relation using existing category triples and knowledge triples in KBs. In the second step of our approach, we mine C2K rules by applying discovered lexical patterns to a hierarchy of categories. In the last step of our approach, we enlarge initial KBs by predicting triples using mined C2K rules. We bootstrap mined rules by repeating the three steps of our approach until there are few rules further mined. In the next, we describe our method in detail.

4.1 The Mining Procedure

In the mining procedure, there are two main conditions to be checked. When the conditions are checked, The prefixes like *dbpedia:*, *dbpedia-owl:* and *Category:* are removed. The first condition is as follows:

- If there are $\langle e, r_{cat}, c\rangle \in T_{cat}$ and $\langle e, r, o\rangle \in T_{know}$ such that $c = o$, then mine the rule:

$$\langle x, r_{cat}, c\rangle \Rightarrow \langle x, r, o\rangle$$

  where x is a variable for entities.

For example, if there are the category triple ⟨*John McCarthy, r_{cat}, Category:Computer scientist*⟩ and the knowledge triple ⟨*John McCarthy, occupation, computer scientist*⟩, then the condition is satisfied and the C2K rule containing the two triples will be made (DBpedia prefixes like *dbpedia*: and *dbpedia-owl:* are omitted for simplicity).

The second condition is as follows:

– If there are $t_{know} = \{\langle e, r_i, o_i \rangle\}_{i=1}^{n} \subseteq T_{know}$ and $t_{cat} = \langle e, r_{cat}, c \rangle \in T_{cat}$ such that all objects of triples in t_{know}, i.e., $\{o_i\}_{i=1}^{n}$, are a substring of $c = (w_1, o_1, w_2, o_2, ..., o_n, w_{n+1})$ and all words in $\{w_2, ..., w_n\}$ have no zero length, then follow the next steps:
  1. Make a co-lexical pattern $p_{co} = (w_1, x_{r_1}, w_2, x_{r_2}, ..., x_{r_n}, w_{n+1})$ for relation $r_1, r_2, ..., r_n$ where $x_{r_{i \in \{1,2,...,n\}}}$ is a variable for an object of r_i.
  2. Make a set of candidate categories which will be compared with p_{co}. We define a set of candidate categories as follows:
  $$C_{candi} = \{c\} \cup siblings(c)$$
  where $siblings(x)$ is a set of siblings of a category x on H_{cat}.
  3. If $c_{candi} = (v_1, v_2, ..., v_m) \in C_{candi}$ is matched with the extracted co-lexical pattern p_{co}, i.e., $v_1 = w_1, v_3 = w_2, ..., v_m = w_{n+1}$, then mine the rules:
  $$\langle x, r_{cat}, c_{candi} \rangle \Rightarrow \langle x, r_1, v_2 \rangle$$
  $$\langle x, r_{cat}, c_{candi} \rangle \Rightarrow \langle x, r_2, v_4 \rangle$$
  $$...$$
  $$\langle x, r_{cat}, c_{candi} \rangle \Rightarrow \langle x, r_n, v_{m-1} \rangle$$
  where x is a variable for entities.

For example, if there are the knowledge triples ⟨*Sohyang, nationality, South Korea*⟩ and ⟨*Sohyang, gender, female*⟩ and the category triple ⟨*Sohyang,* r_{cat}*, Category:female singers of South Korea*⟩, the co-lexical pattern "x singers of y" will be extracted (DBpedia prefixes like *dbpedia*: and *dbpedia-owl:* are omitted for simplicity). Then our method propagates the pattern through siblings like *Category:male singers of England* of the category *Category:female singers of South Korea*. Finally our method can get C2K rules containing not only the category triples with the categories *Category:female singers of South Korea* and *Category:male singers of England*, but also the knowledge triples representing the nationality and gender relations.

Bootstrapping Mined Rules. An initial KB can be sparse, i.e., some entities have many category triples but few knowledge triples. If predicted triples by mined rules are used as a part of knowledge triples, more rules can be mined than the previously mined ones. In this intuition, we bootstrap mined rules through an iterated bootstrapping process. Let Q be a set of rules mined from a KB K and Q^* denotes a set of trustworthy rules which can be defined as rules with a ρ proportion of high confidence rules of Q where ρ is the first parameter of our method (confidence measures will be introduced in the later section). The n-th bootstrapped K can be defined as $K_n = K_{n-1} \cup pred(Q^*_{n-1})$ where $pred(x)$ is a set of new triples predicted by rules contained in a set x and Q^*_{n-1} is trustworthy rules mined from K_{n-1}. Overall iterated bootstrapping process can be represented as follows:

$$K_0 \rightarrow K_1 \rightarrow ... \rightarrow K_n$$

which follows the condition $\frac{|Q_n - Q_{n-1}|}{|Q_{n-1}|} \leq \theta$ where θ is the second parameter of our method which means a threshold of a proportion of increases in mined rules. Q_n is a final output of our method.

4.2 Confidence Measures for C2K rules

Transactions C2K rules are mined on a list of transactions. A transaction is a set of triples in T_{know}, which share the same subject. More precisely, a transaction with n items that share the same entity e can be represented as $T_e = \{\langle e, r_i, o_i \rangle\}_{i=1}^{n} \subseteq T_{know}$.

Support We define a support of a set of triples (sharing an entity e as their subject) as follows:

$$supp(\{\langle e, r_i, o_i \rangle\}_{i=1}^{n}) = \mathbb{I}(\{\langle e, r_i, o_i \rangle\}_{i=1}^{n} \subseteq K)$$

where $\mathbb{I}(x)$ is an indicator function that is the value 1 when a statement x holds, otherwise 0.

A shared subject of triples can be a variable x for an entity. In this case, a support is defined as follows:

$$supp(\{\langle x, r_i, o_i \rangle\}_{i=1}^{n}) = \sum_{e \in E} \mathbb{I}(\{\langle e, r_i, o_i \rangle\}_{i=1}^{n} \subseteq K)$$

Standard Confidence. Confidence of a rule indicates how trustworthy a rule is. A standard confidence of a C2K rule can be defined as follows:

$$conf(t_{cat}^{x} \Rightarrow t_{know}^{x}) = \frac{supp(\{t_{cat}^{x}, t_{know}^{x}\})}{supp(\{t_{cat}^{x}\})}$$

where $t_{cat}^{x} = \langle x, r_{cat}, c \in C \rangle$, $t_{know}^{x} = \langle x, r \in R - \{r_{cat}\}, o \in E \cup L \rangle$, and x is a variable for an entity $e \in E$.

Unnormalized Confidence. The standard confidence of a rule can be abnormally low or high when KBs contain knowledge with some sparsity. In order to discourage abnormal confidence of a rule caused by sparsity of KBs, we calculate a confidence of a rule with only numerator as follows:

$$conf(t_{cat}^{x} \Rightarrow t_{know}^{x}) = supp(\{t_{cat}^{x}, t_{know}^{x}\})$$

where $t_{cat}^{x} = \langle x, r_{cat}, c \in C \rangle$, $t_{know}^{x} = \langle x, r \in R - \{r_{cat}\}, o \in E \cup L \rangle$, and x is a variable for an entity $e \in E$.

In the later experiments, we will show our simple measure is more effect to estimate precision of predictions than those of the standard confidence measure.

5 Experiments

We conducted 3 groups of experiments. In the first group, we compare our approach with AMIE which is the state-of-the-art ARM system publicly available. In the second group of experiments, we analyze the results of our approach more deeply, mostly with regard to the amount and precision. In the last group of experiments, we compare the standard confidence measure to the unnormalized confidence measure introduced in this paper (Section 4).

5.1 Dataset and Experimental Setting

Knowledge Bases. We run our experiments on DBpedia dataset which contains RDF triples extracted from various sources in Wikipedia. Among entire triples with various sources, we only use category-driven triples as category triples (T_{cat}) and infobox-driven triples as knowledge triples (T_{know}). We use English DBpedia 2.0 as an input KB. Because the English dataset is too large to deal with, we randomly choose 20% of triples in English DBpedia 2.0 and only use it in the entire experiments, which will be called $Sample(K_{en})$. $Sample(K_{en})$ contains 1,027,213 category triples and 2,527,466 knowledge triples. We extract a category hierarchy H_{cat} from a Wikipedia category network and use it in the entire experiments.

Settings. Our method have two parameters to be predetermined (Section 4. We set the first parameter ρ as 0.1 and the second parameter θ also as 0.1. AMIE is configured to extract only rules of length 2 which have a category triple as a body of a rule and a knowledge triple as a head of a rule. The other settings of AMIE remain unchanged.

5.2 Results and Analysis

Our Approach vs. AMIE. We run our approach and AMIE on the sampled dataset of DBpedia 2.0, i.e., $Sample(K_{en})$, and mine rules using each method. Our approach mine 150,762 rules with three iterations and AMIE mine 6,959 rules. We predict new triples using the mined rules and check whether new triples are contained in the DBpedia 3.8 dataset. The amount of triples contained in DBpedia 3.8, namely hit count, of each method are measured and compared. Figure 2 shows the results. The left graph shows that our approach predicts more triples which are contained in DBpedia 3.8 than the AMIE's results. This tells us that our approach can mine more useful rules than AMIE's, i.e., the knowledge extracted by rules of our method is more probable in the real world than those of AMIE. The right graph in Figure 2 shows hit proportion of each method, which is a hit count over the amount of entire new triples predicted by each method. The hit proportion would be proportional to the quality of predictions. The right graph shows that our method have high hit proportion than that of AMIE, which indicates that the knowledge extracted by rules of our method is of better quality than that of AMIE. The both graphs show that our method outperforms AMIE in the domain of C2K rule mining. Because of

various features appropriate for C2K rules, which are lexical and hierarchical features of categories, our method can outperform AMIE in the domain of C2K rule mining. The detailed values of results are shown in Table 1 which of rows indicate a hit count over the number of entire new triples predicted by each method with different used proportion of T_{know} in $Sample(K_{en})$.

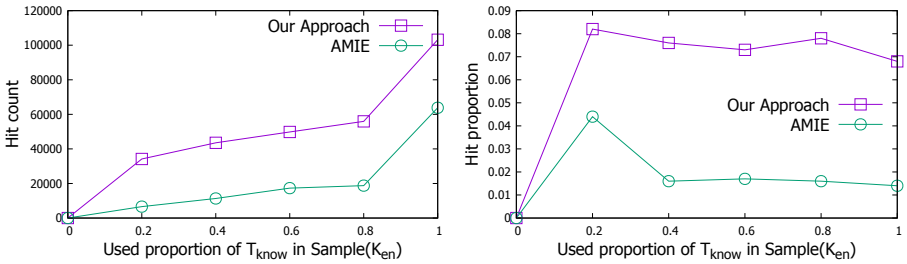

Fig. 2. The hit count (left) and hit proportion (right) of our approach and AMIE with different used proportions of T_{know} in $Sample(K_{en})$

Prop. T_{know}	Our Approach	AMIE
1.0	$103,177_{/1,528,253}$	$63,819_{/4,708,163}$
0.8	$55,945_{/721,611}$	$18,749_{/1,164,978}$
0.6	$49,813_{/678,008}$	$17,308_{/1,047,907}$
0.4	$43,526_{/574,491}$	$11,317_{/714,909}$
0.2	$34,185_{/415,676}$	$6,603_{/150,529}$

Table 1. The hit count and the number of entire new triples predicted by our approach and AMIE with different used proportions of T_{know} in $Sample(K_{en})$

The Amount and Precision of Results. Our approach extracts totally 150,762 rules. The 125,936 rules are initially extracted in the first iteration, and then the rules are bootstrapped to 150,762 rules through three iterations, i.e., 19.71% rules are bootstrapped by iterated bootstrapping. We successfully predict 1,530,253 new triples from 1,027,213 category triples of $Sample(K_{en})$ using the mined rules. Figure 3 and 4 show the examples of the new triples. The entire results can be downloaded at the website [2].

We estimate precision of new triples predicted by the mined rules of our approach. Since there is no computer-processable ground truth of suitable extent [2], we have to rely on manual evaluation. Two people manually evaluate 200 randomly selected samples of the new triples. If a new triple can be inferred from a source category triple, it is regarded as a right one. Domain and range

[2] `http://elvis.kaist.ac.kr/demos/iswc2015_workshop`

are also checked. Only those accepted by both of two people are regarded as true positives.

Table 2 shows an estimated precision of predicted triples which are extracted from category triples of $Sample(K_{en})$ using the mined rules of our approach. We sort the 200 samples by each confidence of rules used to predict them and divide it into 10 equal segments, and then we estimate the precision of the each segment. The table shows that triples predicted by high confidence rules tend to also have high precision, which means that our confidence measure can be used to estimate the precision of predictions. In the later subsection, we will further show the effectiveness of our measure by comparing with the standard confidence measure.

Categories	Subjects	Predicates	Objects	Conf.
Living people	Jennifer Blushi	*dateOfDeath*	living	41185.0
English-language films	Larger Than Life (film)	*language*	English-language	1727.0
People from New York City	Deborah Orin	*birthPlace*	New York City	558.0
Harvard University alumni	William Warren Bartley	*almaMater*	Harvard University	339.0
Liberal Party of Canada MPs	Jacob Thomas Schell	*party*	Liberal Party of Canada	210.0
Queens Park Rangers F.C. players	Dean Wilkins	*clubs*	Queens Park Rangers F.C.	66.0
PlayStation 2 games	Dynasty Tactics series	*ports*	PlayStation 2	0.0

Fig. 3. The examples of true positives

Categories	Subjects	Predicates	Objects	Conf.
People from Philadelphia	Tom Brandi	*location*	Philadelphia	222.0
History of Texas	Wild Cat Bluff, Texas	*placeOfBirth*	Texas	56.0
Athletes at the 1932 Summer Olympics	Babe Zaharias	*medalsilverProperty*	1932 Summer Olympics	24.0
Union College, New York alumni	Orson Spencer	*placeOfDeath*	New York	3.0
National Conference Pro Bowl players	Alfred Jenkins	*nationality*	national	0.0

Fig. 4. The examples of false positives

Unnormalized Conf. vs. Standard Conf. We compare our measure, the unnormalized confidence, with the standard confidence. Figure 5 shows the precision of each segment of each method. The figure shows that the unnormalized confidence values are more close to the precision line of an ideal confidence measure than those of the standard confidence which tend to abnormally high or low in some segments. Table 2 above shows the detailed figures of results. The conclusion of the experiment is that our simple measure is more effective than the standard measure to discriminate importance of C2K rules.

	Unnormalized Confidence		Standard Confidence	
Seg. #	Precision	Avg Conf.	Precision	Avg Conf.
1	0.85	14788.8	0.85	1.0
2	0.7	121.55	0.7	1.0
3	0.45	22.25	0.35	1.0
4	0.55	3.4	0.65	0.34
5	0.55	1.0	0.45	0.06
6	0.6	1.0	0.55	0.014
7	0.8	1.0	0.95	0.0014
8	0.45	0.35	0.45	8.49e-06
9	0.25	0.0	0.25	0.0
10	0.35	0.0	0.35	0.0

Table 2. The precision and the average confidence values of each segment in samples sorted by the unnormalized confidence and the standard confidence values

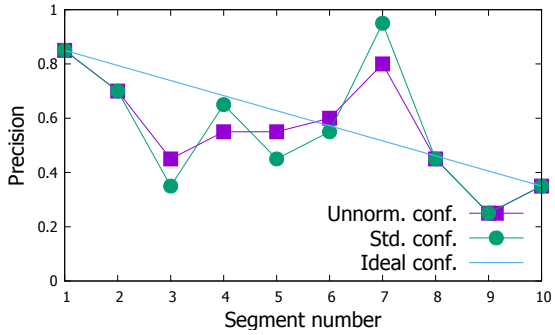

Fig. 5. The precision line of the unnormalized confidence measure, the standard confidence measure and an ideal confidence measure

5.3 Beyond The Current DBpedia

With our approach, DBpedia can be automatically enriched in every time that new category triples come in. The table 3 shows the persistently growing size of category triples in each version of DBpedia. Each column of $|T_{cat}|$ indicates the number of entire category triples in each version of DBpedia. Each column of New $|T_{cat}|$ indicates the number of category triples which does not exist in a immediately previous version of DBpedia (Only New $|T_{cat}|$ of DBpedia 3.8 is the number of category triples not in DBpedia 2.0). With high quality C2K rules mined by our approach, we can persistently construct KBs beyond the current DBpedia from growing Wikipedia categories.

6 Conclusion and Future Work

Throughout this paper, we have proposed an effective method for mining C2K rules. We have also proposed an effective confidence measure to discriminate

DBpedia	3.8	3.9	2014
$\|T_{cat}\|$	15,112,372	16,598,682	18,731,754
New $\|T_{cat}\|$	12,580,437	2,693,767	3,513,358

Table 3. The growing number of category triples in each version of DBpedia.

importance of C2K rules extracted from sparse KBs. Our approach has been proven capable of mining a larger number of C2K rules which can predict more probable and accurate new knowledge from categories than those of the state-of-the-art ARM system for ontological KBs although the systems are for mining more general rules than C2K rules. Our approach mainly utilize lexical and hierarchical features of categories, which is the main reason of our outperformance. Our idea of using these features might be applied on the existing ARM systems to enhance their capability. With our approach, it is possible to persistently and automatically enrich DBpedia-like KBs whose entities are classified in several human-readable categories organized in a network. In the later research, we will further enhance our system and distribute extracted results publicly.

7 Acknowledgement

This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIP) (No. R0101-15-0054, WiseKB: Big data based self-evolving knowledge base and reasoning platform)

References

1. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z. : Dbpedia: A nucleus for a web of open data. pp. 722-735. Springer Berlin Heidelberg (2007)
2. Suchanek, F. M., Kasneci, G., Weikum, G. : Yago: a core of semantic knowledge. In Proceedings of the 16th international conference on World Wide Web, pp. 697–706. ACM (2007)
3. Suchanek, F. M., Kasneci, G., Weikum, G. : Yago: A large ontology from wikipedia and wordnet. Web Semantics: Science, Services and Agents on the World Wide Web, 6(3), 203–217. (2008)
4. Liu, Q., Xu, K., Zhang, L., Wang, H., Yu, Y., Pan, Y. : Catriple: Extracting triples from wikipedia categories. In The Semantic Web, pp. 330-344. Springer Berlin Heidelberg (2008)
5. Nastase, V., Strube, M. : Decoding Wikipedia Categories for Knowledge Acquisition. In AAAI, Vol. 8, pp. 1219-1224. (2008)
6. Galrraga, L. A., Teflioudi, C., Hose, K., Suchanek, F. : Amie: association rule mining under incomplete evidence in ontological knowledge bases. In Proceedings of the 22nd international conference on World Wide Web, pp. 413-422. International World Wide Web Conferences Steering Committee (2013)
7. Dehaspe, L., Toivonen, H. : Discovery of frequent datalog patterns. Data Mining and knowledge discovery, 3(1), 7-36. (1999)

8. Dehaspe, L., Toivonen, H. : Discovery of relational association rules. In Relational data mining, pp. 189-212. Springer Berlin Heidelberg (2001)
9. Goethals, B., Van den Bussche, J. : Relational Association Rules: Getting W armer. Pattern Detection and Discovery. 145-159. (2002)
10. Hearst, M. A. : Automatic acquisition of hyponyms from large text corpora. In Proceedings of the 14th conference on Computational linguistics-Volume 2, pp. 539-545. Association for Computational Linguistics (1992)