*Accepted in the poster track of 5th Joint International Semantic Technology Conference

DBQuote: A Social Web based System for Collecting and Sharing Wisdom Quotes

Guangyuan Piao and John G. Breslin

Insight Centre for Data Analytics
National University of Ireland Galway
IDA Business Park, Lower Dangan, Galway, Ireland
{guangyuan.piao}@insight-centre.org
{john.breslin}@nuigalway.ie

Abstract. Wikiquote, as a project in the Wikipedia family, contains a great number of wisdom quotes and can be consumed by people like journalists for writing. However, statistics show that the number of active editors of Wikiquote is decreasing (-15% for the English Wikiquote and -100% for the Korean Wikiquote). In contrast, Social Web platforms such as Twitter have been increasing in popularity and users are creating more User-Generated-Content (UGC) ever before. In this paper, we present a system named DBQuote, a Wiki and Social Web based system to collect and process wisdom quotes, and then share them with semantic annotation using standard vocabularies such as FOAF, SIOC(T). The system shows that UGC can be used as a source to collect wisdom quotes and that those quotes can be reused instead of providing transient consumption within the Social Web.

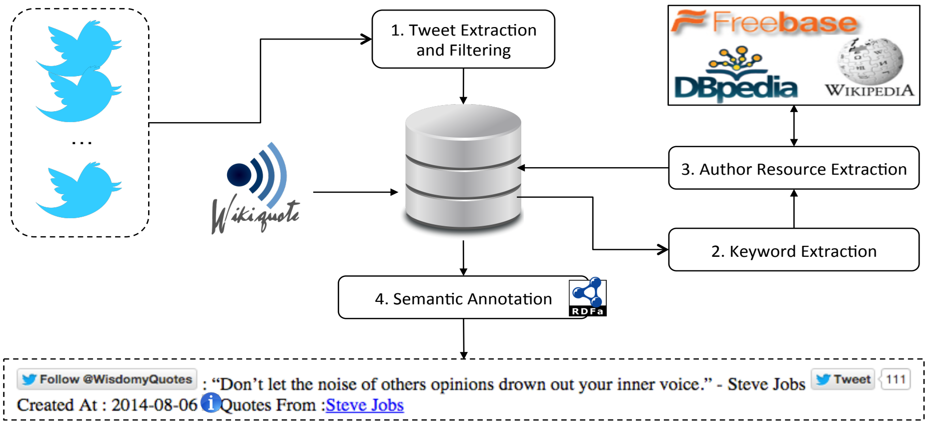
Keywords: Social Semantic Web, DBpedia, Twitter, Wikiquote

1 Introduction

Wikiquote[1] is a free online compendium of sourced quotations from notable people and creative works in every language. People use wisdom quotes in writing since it helps keep perspective, inspire and motivate others. As one might expect, publishing wisdom quotes from Wikiquote in a machine-readable way similar to DBpedia, can benefit both users and application developers. Unfortunately, the statistics of Wikiquote show that the number of active editors of the system has been decreasing, which leads to only 4 new articles per day in English and 0 in Korean[2]. In contrast, users are producing a great amount of User-Generated-Content (UGC) within the Social Web and platforms such as Twitter where the contents cover a lot of topics including wisdom quotes. In this regard, the Social Web can be a valuable source for collecting wisdom quotes in addition to Wikiquote. For instance, there are only 500+ articles in the Korean Wikiquote

[1] https://en.wikiquote.org/wiki/Main_Page

[2] http://stats.wikimedia.org/wikiquote/EN/Sitemap.htm



: "Don't let the noise of others opinions drown out your inner voice." - Steve Jobs
Created At : 2014-08-06 Quotes From :Steve Jobs

Fig. 1. System Workflow of DBQuote

with little new articles (quotes). We can assume that tweets contain more new wisdom quotes. However, there are various challenges to using and processing UGC from the Social Web [**?**], such as *short messages* (e.g., 140 characteristics for tweets) and *noisy content*. In this paper, we investigate if UGC can be used as a source for collecting wisdom quotes that are not available from Wikiquote.

2 System Workflow

First, the system extracts all wisdom quotes with author information by using the Wikiquote API[3]. As a result, 108,846 and 1,655 wisdom quotes were extracted in English and Korean separately. Then the system extracts and processes tweets related to wisdom quotes from some Twitter accounts that have been sharing quotes as per the workflow shown in Fig. 1.

2.1 Tweet extraction and filtering

In order to check the possibility of tweets being a source for collecting wisdom quotes, we randomly selected 10 Twitter accounts that contains "quote" in their account names (5 in English and 5 in Korean) and that have been publishing wisdom quotes on Twitter. By using the Twitter API[4], we extracted 1,000 tweets from each account. Due to the characteristics of Twitter (following and retweets), there are many duplicated items among these accounts. This means that there are many tweets that have been tweeted more than one time since many of these accounts are following each other and retweet the contents of others as well. Thus, we filter out similar tweets using a similarity function in PHP. The extracted elements from tweets for the semantic annotation process are: (1) user

[3] `https://github.com/natetyler/wikiquotes-api`
[4] `https://dev.twitter.com/rest/public`

account - extracts original creator if it was retweeted, (2) tweet content, (3) created time and (4) hashtags (if they exist).

2.2 Keyword extraction

To extract resources related to the authors of quotes, we use the Zemanta API[5], which can extract personal information and provide related links from knowledge bases such as Freebase for a given sentence. It has great functionality, but not without limitations. In order to extract additional author information within quotes, we use an n-gram (n = 3, 2, 1) approach for each wisdom quote, extract a keyword list, and then use the list to query DBpedia using SPARQL [?].

2.3 Resource extraction for author information

For English wisdom quotes, 1,356 resources of type `Person` were extracted using the Zemanta API from the collected 3,679 tweets. In order to extract author information which cannot be retrieved by the Zamanta API, we look up resources in DBpedia via its SPARQL endpoint. Extracting the author's name for a wisdom quote is not a trivial task. We use the keyword list to look up resources of type `Person` or a subclass of `Person` in DBpedia. For the Korean DBpedia, due to the lack of APIs such as Zemanta for retrieving personal information from open knowledge bases, we only use the keyword list to look up resources in Korean DBpedia. We look up a resource of type `Person` or one that has the property `dbpedia-owl:birthDate`[6] even it is not represented as a type of `Person`. After all, 2,258 author names were extracted out of 3,679 English tweets and 325 author names were extracted out of 1,473 Korean tweets.

2.4 Semantic annotation of wisdom quotes

It is important to share content in a way so that a person or machine can explore the data according to Linked Data Principles[7]. While publishing posts in DBQuote, we use RDFa[8] for semantic annotation. By doing so, all wisdom quotes are machine readable as part of the Web of Data. There are several ontologies or vocabularies for representing microblogs. We adopt FOAF [?], SIOC [?] and the extensions (SIOCT) and CiTO ontology[9]. These vocabularies cover all of the information we want to represent within our system and the ontology that is used for semantic annotation of posts in DBQuote is displayed in Fig. 2.

[5] `http://www.zemanta.com/api/`

[6] The prefix `dbpedia-owl` denotes the namespace `http://dbpedia.org/ontology/`

[7] `http://www.w3.org/DesignIssues/LinkedData.html`

[8] `http://www.w3.org/TR/xhtml-rdfa-primer/`

[9] `http://purl.org/spar/cito`

Table 1. Statistics of collected quotes from Wikiquote and 10 Twitter accounts

	Total	**From Wikiquote**	**From Twitter**
# of English quotes	110,992	108,846	2,146
# of Korean quotes	1,976	1,655	321

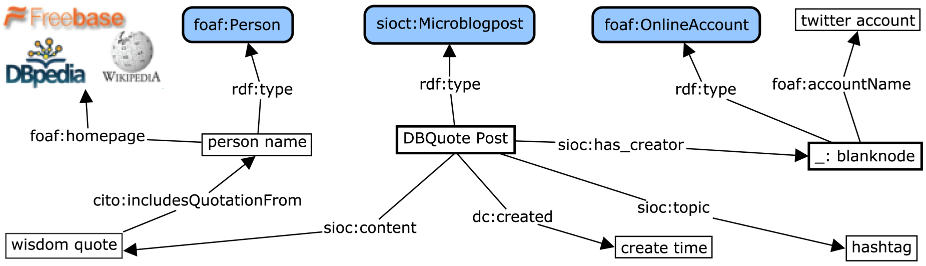


Fig. 2. Ontology for semantic annotation in DBQuote

3 Conclusions

The statistics of collected quotes from the Wikiquote and 10 Twitter accounts are displayed in Table 1. 2,146 (+2%) and 321 (+19%) additional quotes were produced by collecting 1,000 tweets from each account. It shows UGC has a great potential to be a source for collecting wisdom quotes that complement quotes from Wikiquote especially for languages with a lack of contributors (e.g., the Korean Wikiquote). In the future, we plan to check latent relationships among wisdom quotes. For example, for those wisdom quotes with more than one author, we can use detailed information from knowledge bases such as DBpedia (e.g., `dbpedia-owl:birthDate`(s)/`dbpedia-owl:deathDate`(s) for authors) to identify the provenance.

Acknowledgments. This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289 (Insight Centre for Data Analytics).

References

1. Bontcheva, K., Rout, D.: Making sense of social media streams through semantics: A survey. Semantic Web 5(5), 373–403 (2014)
2. Breslin, J.G., Harth, A., Bojars, U., Decker, S.: Towards semantically-interlinked online communities. In: The Semantic Web: Research and Applications, pp. 500–514. Springer (2005)
3. Brickley, D., Miller, L.: FOAF vocabulary specification 0.98. Namespace document 9 (2012)
4. PrudHommeaux, E., Seaborne, A.: SPARQL query language for RDF. W3C recommendation 15 (2008)