Automatically Predicting Quiz Difficulty Level Using Similarity Measures

*

Chenghua Lin
Computing Science
University of Aberdeen
Aberdeen, AB24 3UE, UK
chenghua.lin@abdn.ac.uk

Dong Liu
BBC Future Media &
Technology
Knowledge & Learning
Salford, M50 2QH, UK
Dong.Liu@bbc.co.uk

Wei Pang
Computing Science
University of Aberdeen
Aberdeen, AB24 3UE, UK
pang.wei@abdn.ac.uk

Edward Apeh
Computing Science
Bournemouth University
Poole, BH12 5BB, UK
eapeh@bournemouth.ac.uk

ABSTRACT

In this paper, we present a semi-automatic system (Sherlock) for quiz generation using Linked Data and textual descriptions of RDF resources. Sherlock is distinguished from existing quiz generation systems in its ability to control the difficulty level of the generated quizzes. We cast the problem of perceiving the level of knowledge difficulty as a similarity measure problem and propose a novel hybrid semantic similarity measure using linked data. Extensive experiments show that the proposed similarity measure outperforms four strong baselines in both the pilot evaluation using a synthetic gold standard as well as with human evaluation, giving more than 47% gain in clustering accuracy over the baselines.

Categories and Subject Descriptors

I.2 [**Artificial Intelligence**]: Miscellaneous

General Terms

Algorithms, Experimentation

Keywords

Quiz Generation, Linked Data, RDF, Educational Games, Semantic Similarity, Text Analytics

1. INTRODUCTION

Interactive games are effective ways of helping knowledge being transferred between humans and machines. On one hand, efforts have been made to design games with the purpose of semi-automating a wide range of knowledge transfer tasks by leveraging the wisdom of the crowd. For instance, symmetric and asymmetric verification games [15, 8] have been developed for assisting with Semantic Web tasks such as ontology building, ontology alignment, content annotation and entity interlinking. Likewise, some researchers have developed quiz like games to rank, rate and clean up Linked Data [19, 18]. In this way, the factual knowledge is transferred from humans, especially domain experts, to computers. On the other hand, work has also been done to unleash the potential of Linked Data in generating educational quizzes for aiding learners' knowledge acquisition from a knowledge base. When building a quiz generation system using Linked Data, most existing approaches are based on domain specific templates and the creation of quiz templates heavily relies on the ontologists and Linked Data experts [4, 1].

Moreover, a system that can generate quizzes with different difficulty levels will better serve users' needs. However, such an important feature is rarely offered by the existing systems. Some researchers [18] determined the difficulty of a quiz by simply assessing the popularity of a RDF resource, without considering the fact that the difficulty level of a quiz is directly affected by the selection of wrong answer candidates. Also, the most common way of generating wrong answer candidates is to randomly select from the results of querying Linked Data repositories and hence provides no means of controlling the difficulty level in quiz generation. To our knowledge, while different similarity measures have been widely used for measuring the degree of closeness or separation of target objects, the problem of how well similarity measures can be used to represent the degree of knowledge difficulty in terms of human perception is still relatively unexplored.

In this paper, we demonstrate a novel semi-automatic quiz genera-

K-CAP 2015, October 07 - 10, 2015, Palisades, NY, USA. Copyright is held by the owner/author(s). Publication rights licensed to ACM. ACM 978-1-4503-3849-3/15/10$15.00
DOI: http://dx.doi.org/10.1145/2815833.2815842

tion system (Sherlock) empowered by semantic and machine learning technologies. Sherlock is distinguished from existing quiz generation systems by its ability to control the difficulty level of the generated quizzes, based on a novel Linked Open Data (LD) based hybrid semantic similarity measure, called TF-IDF (LD).

For investigating how well the proposed algorithm can be used to represent the difficulty level of knowledge (i.e. *difficult*, *medium* and *easy*), we evaluated the performance of TF-IDF (LD) against four strong baselines (i.e., two knowledge-based and two text-based similarity measures) based on the BBC Wildlife dataset[1]. It was observed that the knowledge-based measures give better performance in predicting the *easy* cluster compared to the text-based measures, but are inferior in the prediction for the *difficult* and *medium* clusters. Our proposed hybrid semantic measure TF-IDF (LD) outperforms four strong baselines and gives at least 50% gain in overall clustering accuracy. In addition, it was discovered in the human evaluation that model accuracy derived from human quiz test indeed showed a strong correlation with the pair-wise quiz similarity, with the proposed algorithm again giving the best performance.

The rest of the paper is organised as follows. Section 2 and 3 reviews the related work and presents the Sherlock architecture, respectively. The hybrid semantic similarity algorithm is detailed in Sections 4. Finally, experimental results are discussed in Section 5 and we conclude the paper in Section 6.

2. RELATED WORK

2.1 Games with a Purpose and Educational Games

A series of symmetric and asymmetric verification games were presented in [15] with the aim to motivate humans to contribute to building the Semantic Web. BetterRelations [8] is a representative symmetric verification game built following the concepts of "Games with a purpose", which attempts to solve the problem of ranking RDF triples within the description of an entity. Other quiz-like games focus on ranking, rating and cleansing Linked Data [18, 19] . The assumption underlying these games is that the frequency of a question being correctly answered implies the importance of the supporting Linked Data used to create the quiz. However, the focus of these games is to harness human intelligence to perform tasks that cannot be automated, rather than creating learning experiences for humans.

In contrast to games with a purpose, Damljanovic et al. presented a template-based method for generating educational quizzes, in which a conversational AI agent was introduced to guide the learners and to dynamically select quizzes according to the learners' needs [4]. Linked Data Movie Quiz (LDMQ) is another representative work of using Linked Data for template-based quiz generation, which is able to generate quizzes related to a user-selected actor or actress, asking questions about the director, release date or the characters of a film in which the actor or actress appeared [14].

One of the common limitations shared by existing quiz generation systems is the domain dependent issue. That is whenever applying the template-based quiz generation method to a new domain, significant human efforts will be required on tasks such as creating new question templates, writing SPARQL queries according to a domain-specific ontology and defining rules for collecting wrong answers for a quiz. In addition, the difficulty of quizzes plays an important role in formal learning. However, as stated in [18], many quiz generation systems suffer from the fact that the generated quizzes can be either "too simple or too difficult", largely due to the lack of quantitative analysis on the relation between the wrong answer candidates and the correct answer. This has in turn motivated us to develop a systematic way of measuring quiz difficulty level using semantic similarity measures.

2.2 Similarity Measures

2.2.1 Text-based Approaches

Measures of text similarity have been used for a long time in applications in natural language processing and related areas. The text-based measures try to identify the degree of the similarity between text units using statistical patterns of words derived from large corpora, where the most representative measures are cosine similarity, averaged Kullback-Leibler divergence (KLD) and the squared Euclidean distance [9]. Some more advanced approaches rely on word-co-occurrence patterns derived from large corpus which indicate the degree of statistical dependence between text units, and the statistical dependences can then be used as a measure for text similarity. Further representative approaches in this line include pointwise mutual information (PMI) [17], latent semantic analysis (LSA) [10], and models that belong to the topic model families [2, 12].

2.2.2 Knowledge-based Approaches

In contrast to corpus-based approaches which are purely oriented on statistical techniques, knowledge-based approaches relied on human-organised knowledge (e.g. semantic Network, WordNet and Linked Open Data) to encode relations between a collection of concepts [6, 7, 13].

WordNet is a large English lexical knowledge database in which terms are grouped into different sets known as synsets with a list of synonyms. A number of measures have been developed based on the WordNet hierarchy such as accessing the semantic relatedness of words/entities [20] and identifying word sense under different context [11]. One of the closest work to our proposed hybrid semantic similarity measure is the Linked Data Semantic Distance (LDSD) [13], which also uses the graph information in RDF resources for computing semantic similarity. However, it should be noted that the similarity computation of LDSD purely rely on the statistics of the direct and indirect in and out connections among RDF resources of DBpedia, without considering the importance of predicates. In contrast to LDSD which can only work with RDF data, the proposed TF-IDF (LD) algorithm can deal with both literal values and textual descriptions.

3. THE SHERLOCK ARCHITECTURE

We first give a brief overview of the Sherlock system as it deploys the proposed TF-IDF (LD) algorithm for controlling the difficulty level in quiz generation and serves as the basis for our experiments. Figure 1 depicts an overview of the Sherlock quiz generation framework, in which the components are logically divided into two groups: online and offline. Within the Sherlock framework, different components can interact with each other via two shared databases that respectively contain information about: (1) questions and answers of quizzes and (2) distractors, i.e., incorrect answers.

Data Collection and Integration We collected two different types of data, which are structured RDF data published by DBpedia and

[1] `http://www.bbc.co.uk/nature/wildlife/`

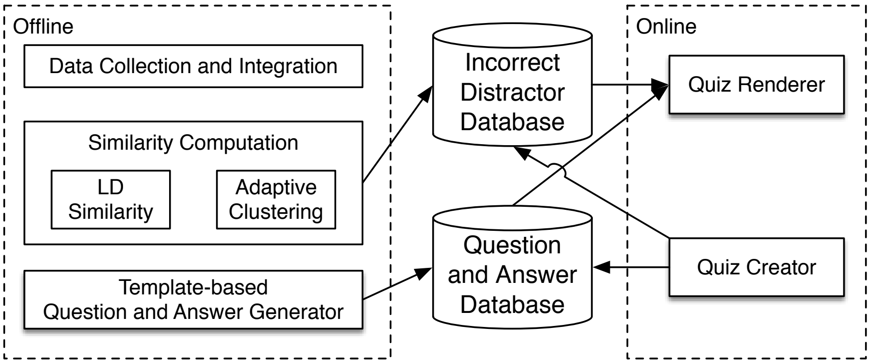

Offline
Data Collection and Integration
Similarity Computation
LD Similarity
Adaptive Clustering
Template-based Question and Answer Generator
Incorrect Distractor Database
Question and Answer Database
Online
Quiz Renderer
Quiz Creator

Figure 1: Overall architecture of Sherlock.

the BBC; and unstructured text describing objects/entities collected from the BBC website and Wikipedia. These datasets play two main roles, i.e., serving as the knowledge base for quiz generation and for calculating the similarity scores between objects/entities. Detailed descriptions on dataset preparation are given in the Experiment section.

Similarity Computation The similarity computation module is the core of the offline part of the architecture. The similarity computation module first accesses the RDF store and the text corpus, and then calculates the similarity scores between each object/entity pair. In the second step, the module performs K-means clustering to partition the wrong candidate answers into different difficulty levels according to their similarity score with the correct answer of a quiz. Here we empirically set K=3 corresponding to three predefined difficulty levels, i.e. "easy", "medium" and "difficult".

Template-based Question and Answer Generator The quiz generator component adopts a template-based method similar to LDMQ [14], which is able to boost the system in the situation of cold-start or coping with data from a new domain. For instance, a template "*Which of the following animals is* `{?animal_name}`?" can be instantiated by replacing the variable with `rdfs:label` of a animal.

Quiz Renderer The quiz renderer module realises the user interface through which users can interact with the system (See Figure 2). The question and correct answer are retrieved from a dedicated database, and the wrong answers candidates are selected from the results of the similarity computation module. It is worth noting that the values of `foaf:depiction` in the RDF store provide the links to the images used for rendering the quiz. The user interface also offers a nice feature by allowing a user to tune up (or down) the difficulty level of the next quiz, depending on whether a user fails or succeeds in playing the current quiz.

4. METHODOLOGY

In this section, we describe the main algorithm we have developed in Sherlock. As we recall, one of the key challenges in our work is to measure the difficulty level of quizzes. To this end, we developed a hybrid similarity measure by combining a novel linked data based TF-IDF scheme with the classical text-based cosine similarity measure, called TF-IDF (LD).

Typically, RDF datasets are formalised as graphs, and the direct and indirect distances in those graphs can be used to measure the similarity between RDF resources, as in the case of Linked Data Semantic Distance (LDSD) [13]. While LDSD is reported to be effective on large scale datasets such as DBpedia and Freebase, the importance of predicates in RDF resources are not considered which limits the accuracy of LDSD. To address this issue, we propose a novel Linked Data based TF-IDF scheme by mapping Named Graphs into vectors, which takes the predicate information into account. The resulting Linked Data based TF-IDF vectors are then combined with the cosine similarity measure for calculating the semantic similarity between two RDF resources. Before describing the proposed algorithm, we first give formal definitions to a few technical terms including: *term*, *sentence*, *document* and *corpus*.

Definition 1: A term and a sentence

A RDF statement, i.e. a tuple of (subject, predicate, object), is defined as a *sentence*. Two types of combinations such as (subject, predicate) and (predicate, object) are regarded as *terms*. For example, `(_:Cheetah, wo:family, _:Felidae)` is a sentence, whereas `(_:Cheetah, wo:family)` and `(wo:family, _:Felidae)` are two terms in the sentence. Here `wo` is the the namespace of BBC Wildlife Ontology[2].

Definition 2: A document and a corpus

A Named Graph that is related to a RDF resource is a *document*, and a collection of documents is a *corpus*.

For example, the RDF statements shown in Listing 1 constitute a document. This document contains three sentences describing the animal cheetah.

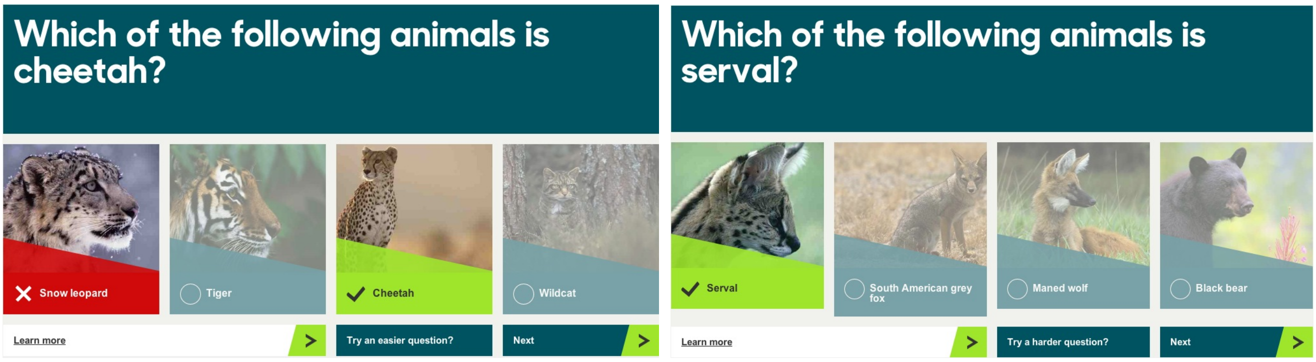
Listing 1: Example of a RDF document.

```
_:Cheetah wo:family _:Felidae ;
  wo:order _:Carnivora ;
  wo:class _:Mammal .

<http://www.bbc.co.uk/news/science-environment
    -22861142> a foaf:Document ;
  foaf:primaryTopic _:Cheetah .
```

Definition 3: The relation between a term, a sentence and a document

[2] http://www.bbc.co.uk/ontologies/wildlife/2010-02-22.shtml



Which of the following animals is cheetah?

(a) User interface when an incorrect choice is made.

Which of the following animals is serval?

(b) User interface when a correct choice is made.

Figure 2: User interface for playing quizzes.

```
Algorithm 1 The TF-IDF (LD) algorithm.
Require: A Corpus $C$ (i.e., a collection of named graphs), named
    graph $a \in C$, named graph $b \in C$, $idf$ value of the terms in $C$
    computed using Equation 2.
Ensure: Semantic similarity between $a$ and $b$.
for each term $(p, o) \in C$ do
    if term $(p, o)$ exist in $\{$a, b$\}$ then
        $w_p = tfidf(p, o) = idf(p, o)$
    else
        $w_p = tfidf(p, o) = 0$
    end if
end for
Derive Linked Data based TF-IDF vectors $\vec{t_a} = \{w_1^a, w_2^a, ..., w_p^a\}$ for $a$; $\vec{t_b} = \{w_1^b, w_2^b, ..., w_p^b\}$ for $b$
Compute semantic similarity $\text{SIM}_C(\vec{t_a}, \vec{t_b}) = \frac{\vec{t_a} \cdot \vec{t_b}}{||\vec{t_a}|| ||\vec{t_b}||}$.
```

If a document d contains a sentence (s, p, o), then we say terms (s, p) and (p, o) are *in* document d, i.e. $(s, p) \in d$ and $(p, o) \in d$.

The TF-IDF(LD) Algorithm We now describe our TFIDF(LD) algorithm. With the definitions above, the classical TF-IDF scheme can then be applied on the RDF datasets. That is given a term t, a document d and a corpus C, the *Term Frequency (TF)* and the *Inverse Document Frequency (IDF)* are calculated as follows:

$$tf(t, d) = \begin{cases} 1 & \text{if } t \in d \\ 0 & \text{if } t \notin d \end{cases} \tag{1}$$

$$idf(t, C) = \log \frac{|C|}{|\{d \in C : t \in d\}|} \tag{2}$$

It is worth noting that in Equation 1 we apply a Boolean TF function for the term frequency calculation, which means that if a triple belongs to a graph, its term frequency is 1 and 0 otherwise. By applying Equations 1 and 2 onto the RDF datasets, Named Graphs can then be transformed into vectors. Once the Linked Data based TF-IDF vectors (i.e., $\vec{t_a}$ and $\vec{t_b}$,) for the two RDF resources have been computed, we can then calculate their semantic similarity using the cosine similarity (Equation 3). A summary of the TF-IDF (LD) algorithm is given in Algorithm 1.

Table 1: Wildlife textual dataset statistics. Note: †denotes before preprocessing and * denotes after preprocessing.

Dataset	# of docs	Avg. doc length†	Avg. doc length*	Vocab. size†	Vocab. size*
Wildlife	437	1,190	652	26,004	18,237

$$\text{SIM}_C(\vec{t_a}, \vec{t_b}) = \frac{\vec{t_a} \cdot \vec{t_b}}{||\vec{t_a}|| ||\vec{t_b}||}. \tag{3}$$

5. EXPERIMENT

In this section, we explore how well similarity measures can be used to suggest quiz difficulty level that matches human's perception, in which we compared several well known similarity measures against the proposed TF-IDF (LD) algorithm. In particular, we have investigated the following research questions:

- Are similarity measure(s) appropriate for assessing quiz difficulty level?
- To what extend can the quiz difficulty level suggested by similarity measure(s) match humans' perception of knowledge difficulty level?

5.1 A Pilot Evaluation on Predicting Quiz Difficulty

We shall not try to give a general definition of difficulty covering a wide range of psychological aspects from emotional problems to intellectual and physical challenges. Instead, we consider the notion of difficulty in the sense used in quiz generation, one that builds as combinations of predefined candidates. Of course the study of the overall difficulty for a given quiz involves multiple factors such as the intellectual level of knowledge covered in the quiz and the user's knowledge background, etc. In the preliminary study, we try to address the problem in a less complicated scenario, in which the difficulty level of a quiz is directly driven by the similarity between the correct answer and the wrong answer candidates.

5.1.1 Data

We conducted the preliminary experiment for measuring quiz difficulty level based on the BBC Wildlife dataset. The choice of

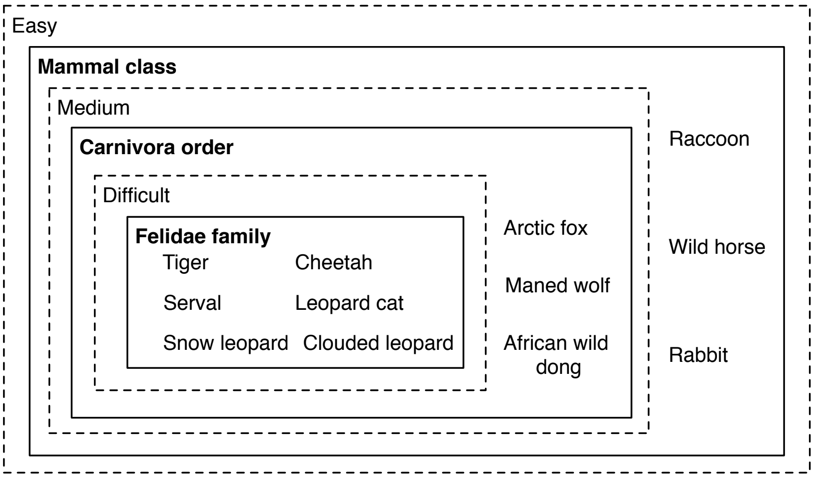


Figure 3: Deriving gold standard for the BBC Wildlife dataset using the biological classification system.

Table 2: Clustering accuracy of different similarity measures for measuring quiz difficulty level. (Unit in %, numbers in bold face denote the best result in their respective row.)

	LDSD	WUP	KLD	TF-IDF	TF-IDF (LD)
Dataset	RDF	RDF	Text	Text	RDF
Difficult	18.4	2.4	37.5	29.2	**85.7**
Medium	7.9	9.3	11.4	11.6	**66.2**
Easy	82	74.5	50.9	44.8	**99.3**
Overall	36.1	28.7	33.3	28.5	**83.7**

dataset for evaluation is based on the fact that (1) there is no readily available gold standard for benchmarking from the literature; (2) in the Wildlife dataset, each animal has been labelled using the biological classification system (i.e., family, order and class), which can be naturally used as gold standard for evaluation; and (3) according to the statistics from the BBC, the BBC Wildlife website is one of the most frequently visited BBC website, indicating a broad public interest in the Wildlife data. In particular, we have prepared two different versions for the BBC Wildlife dataset, i.e., structured RDF data and unstructured textual data.

RDF data: As for the Wildlife dataset, DBpedia and the BBC Wildlife website have already published the RDF data, so we harvested the RDF data directly from these two data sources. In total there are 49,897 RDF triples in the dataset.

Textual data: In addition to the RDF data, we have also collected a dataset containing textual descriptions for objects/entities (e.g. different animals, etc.) from the BBC website and Wikipedia. Here the textual datasets are mainly used for calculating the text-based similarity scores between objects/entities and so are used to predict the quiz difficulty level. In the preprocessing, we used an HTML parser to extract contents from the HTML pages and then removed wildcards and word tokens with non-alphanumeric characters, followed by stop word removal and Porter stemming. The textual dataset statistics are summarised in Table 1.

5.1.2 Experimental Results

To tackle the first research question, in the pilot evaluation, we formulate the problem of perceiving the level of knowledge difficulty as a similarity measure problem. The hypothesis is that if some objects/entities share a lot of (semantically) similar properties, they tend to have higher degree of semantic relatedness with subtle difference and hence more difficult to be disambiguated, and vice versa.

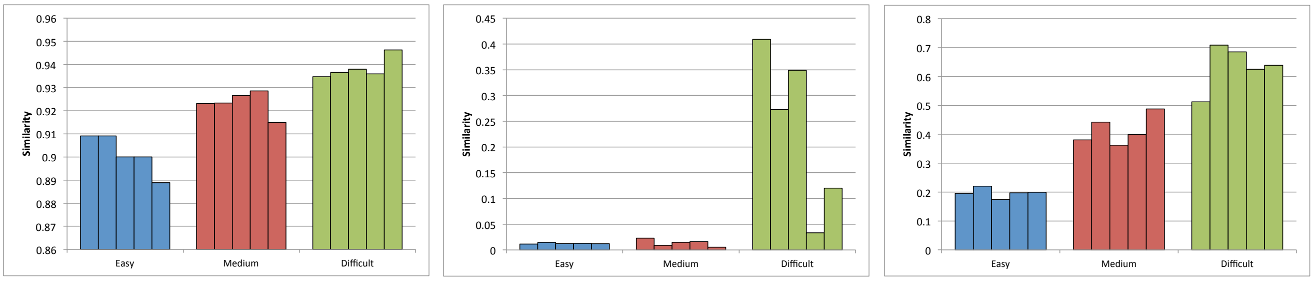
To derive the gold standard for the Wildlife dataset, one intuitive approach is to make use of the biological classification system. We define that if some animals have the same *family* label (e.g., Cheetah and Serval), then these animals will be very similar to each other and hence *difficult* to disambiguate. Likewise, if some animals have the same *order* label but from different *families*, they will be less similar and correspond to a *medium* difficulty level when generating a quiz. Similarly, a quiz generated based on animals with the same *class* label but different *family* and *order* labels will be the most dissimilar and correspond to the *easy* level. Note that the gold standard is derived independently, which does not relied on any information/features used by the baselines or the TF-IDF(LD) algorithm. An illustrative example of the gold standard is shown in Figure 3.

In the pilot evaluation, we compared the proposed hybrid semantic similarity algorithm TF-IDF (LD) against four strong baselines in the task of predicting quiz difficulty level: two knowledge-based similarity measures using the RDF dataset (i.e., LDSD [13] and a WordNet based measure denoted as WUP [20]); and two text-based measures using the textual dataset (i.e., cosine similarity with traditional TF-IDF and KLD).

As can be observed from Table 2 that for the text-based similarity measure, KLD outperforms TF-IDF in predicting the *difficult* and *easy* clusters, whilst having a similar performance for the *medium* cluster. The knowledge-based measures slightly outperform the text-based measure for about 3% in overall. It was also found that compared to the text-based measures, the knowledge-based measures (i.e., LDSD and WUP) give much better performance in predicting the *easy* cluster (i.e. 30% higher), but are inferior to the *difficult* and *medium* clusters prediction. Our TF-IDF(LD) algorithm outperforms all four strong baselines for the three classes,

WUP	KLD	TF-IDF	LDSD	TF-IDF (LD)
Jaguar	Leopard	Leopard	Lion	Serval
Lion	Lion	**Blackbuck**	**Stoat**	Snow Leopard
Serval	Cougar	Lion	Leopard	Lion
Cougar	Tiger	Leopard Cat	Tiger	Leopard
Meerkat	Jaguar	Cougar	Serval	Cougar
Aardvark	**Spotted Hyena**	Asian Golden Cat	Cougar	Wildcat
Coyote	Leopard Cat	**Grant's gazelle**	**Gray Wolf**	Jaguar
Capybara	Snow Leopard	**Spotted Hyena**	**Red Fox**	Tiger
Stoat	**Bongo (antelope)**	**Blue Wildebeest**	**Meerkat**	**Aardvark**
Indri	**Fossa**	Snow Leopard	**Human**	Eurasian Lynx

Figure 4: Top 10 animals similar to Cheetah found by different algorithms (inappropriate ones are highlighted in bold).



(a) LDSD

(b) TF-IDF

(c) TF-IDF (LD)

Figure 5: Averaged quiz similarity based on different similarity measures on the Wildlife domain dataset.

with over 47% improvement in overall accuracy. This demonstrates the effectiveness of the proposed algorithm. One reason why TF-IDF (LD) significantly outperforms the baselines is probably due to the fact that our algorithm not only considers the direct and indirect connections between RDF resources, but also models the information of predicates by treating RDF graphs as documents. In this way our proposed algorithm can capture richer semantic information from data even under the condition where the links between RDF resources are sparse.

To better compare and illustrate the clustering performance, Figure 4 lists the top ten animals that are most similar to Cheetah, grouped by different similarity algorithms. In the figure, animals are listed in descending order based on their similarity to Cheetah, and the ones that are not in the same *family* as Cheetah are highlighted in bold. It can be observed from Figure 4 that, among the four baselines, KLD performs best with three animals in the cluster not belonging to the same family as Cheetah; in contrast, WUP is the least accurate which includes six outliers in the cluster. The TF-IDF (LD) algorithm again gives the best performance with only one outlier *Aardvark* being included.

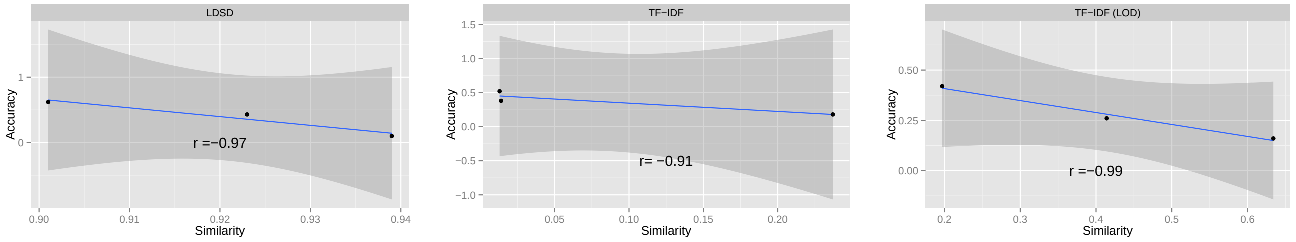
5.2 Using Human Judgements to Examine the Quiz Difficulty Level

Although the previous pilot experiment shows that similarity measures, especially the proposed TF-IDF (LD) algorithm, are potentially appropriate means for measuring quiz difficulty level, this study is still based on a synthetic gold standard without human evaluation. Therefore, it is necessary to verify whether the difficulty level captured by similarity measures is indeed in line with human judgements, and if so how strong the correlations are.

5.2.1 Task Description of Human Evaluation

To investigate the second research question, we propose a task that creates a formal setting for assessing how human perceive knowledge difficulty level, called the quiz game task. Basically the task involves playing quiz games, in which the subject is presented with quizzes produced using three selected similarity measures including LDSD, TF-IDF and TF-IDF(LD), with 5 quizzes generated for each difficulty level per measure. Therefore there are altogether 45 quizzes generated based on the Wildlife dataset using the three different similarity measures. The rationale for using a subset of the baselines are mainly based on the concerns that (1) those four baselines performed quite similar in the pilot study; and (2) more importantly, four baselines plus the proposed algorithm will involved 75 testing quizzes, requiring more than 15 minutes to complete the test. It was reported by [16] that human evaluation test taking more than 15 minutes will result in participants being less focused and more likely to be interrupted.

The above described tasks were offered on Amazon Mechanical Turk, which has been successfully used in the past to develop gold-standard data for various tasks, e.g., natural language processing [3] and labelling images [5]. We presented each subject with jobs containing 45 quiz tasks. Each job (i.e., a quiz) was performed by 30 separate subjects and we recorded the answers picked by the subjects. Also, to reduce the randomness of human evaluation, the subjects are instructed to choose an additional option '*I don't know*', if he/she is not sure about the answer of a quiz. Such a selection will be automatically categorised as an incorrect answer. Completing



(a) LDSD ($r = -0.97, p = 0.156$) (b) TF-IDF ($r = -0.91, p = 0.266$) (c) TF-IDF (LD) ($r = -0.99, p = 0.031$)

Figure 6: Pearson's correlation between the model accuracy and the pair-wise similarity of quizzes.

the whole test takes 12 mins for a subject on average.

5.2.2 Model Accuracy

To quantify the difficulty levels perceived by users in the human evaluation task, we introduced the concept of *model accuracy*, which indicates the percentage of times users have chosen the correct answer of the quizzes generated by a model. Here the model refers to a particular similarity measure (e.g. LDSD or TF-IDF). Let q_k^s be the answer selected by the sth subject for the kth quiz; c_k be the correct answer for the kth quiz, and S denotes the number of subjects, the accuracy of the kth quiz is calculated as follows:

$$\mathrm{A}_k = \sum_s (q_k^s = c_k)/S. \tag{4}$$

Finally, we are interested in calculating the model accuracy M_m^l, which encodes the percentage of times users have chosen the correct answer for the test quizzes of difficulty level l generated by model m. The derivation of M_m^l is formalised in Equation (5)

$$\mathrm{M}_m^l = \sum_k A_k/D, \tag{5}$$

where D is the total number of quizzes with difficulty level l generated by model m.

5.2.3 Correlation between Model Accuracy and Similarity Distribution

In this experiment, we investigated the correlation between the difficulty level suggested by similarity measures and that perceived by human as encoded in the *model accuracy*. Our hypothesis is that if the difficulty levels suggested by similarity measures are in line with human perception, the *pairwise similarity* of the quizzes should have some correlations with the *model accuracy*. Here the *averaged pairwise similarity* of a quiz is calculated by averaging out the similarity scores between the correct answer and distractors of that quiz.

Figure 5 shows the pair-wise quiz similarities of the testing quizzes, where each histogram corresponds to a quiz. It can be observed that the similarities based on LDSD are quite concentrated, with less than 0.06 difference between the quizzes with the highest and lowest similarity values. This is likely due to the fact that LDSD relies on the direct and indirect links between resources, which are relatively sparse in the Wildlife RDF dataset. As a result, LDSD produces similarity values with very small difference for the quizzes. On the other hand, the classic TF-IDF scheme produces a quite skewed similarity distribution, with the *easy* and *medium* clusters having very small similarity values and the *difficult* class having much higher similar scores. In contrast, the quiz similarity distributions for each difficulty level obtained using the proposed TF-IDF (LD) algorithm are much more balanced and well spread.

Figure 6 shows the Pearson's correlation between the model accuracy and the pair-wise similarity of quizzes generated from the same model, in which all the data points are averaged over 5 quizzes per difficulty level. It can be seen that for all the three tested models, model accuracy derived from human evaluation indeed shows a negative correlation with the pair-wise quiz similarity. In addition, the proposed TF-IDF (LD) shows better correlation than both LDSD and TF-IDF in terms of r value. Furthermore, for the significance test, TF-IDF (LD) is the only measure with $p < 0.05$, while the other compared approaches, i.e., LDSD with $p = 0.156$ and TFIDF with $p = 0.266$. The human evaluation results are in line with the observation in the pilot study using the gold standard derived based on the biological classification system. Therefore, we conclude that similarity measures are appropriate means for measuring quiz difficulty level and that the proposed TF-IDF (LD) algorithm is superior to the baselines for predicting quiz difficulty level.

6. CONCLUSION

In this paper, we presented a generic framework (Sherlock) for generating educational quizzes using Linked Data. Compared to existing systems, Sherlock offers a distinctive capability in controlling the difficulty level of the generated quizzes based on a novel hybrid semantic similarity measure TF-IDF (LD). The TF-IDF (LD) algorithm is simple but very effective, which outperforms four strong baselines in both the pilot evaluation using a synthetic gold standard as well as in the human evaluation, giving more than 47% gain in overall clustering accuracy over the baselines.

As for future work, we first plan to carry out more comprehensive user testing and evaluation to further explore the relation between quiz difficulty and semantic similarity. Making the Sherlock system context aware is also an interesting direction as semantic similarity can be context dependent.

7. ACKNOWLEDGMENTS

This work is supported by the BBC Connected Studio programme[3], the award made by the RCUK Digital Economy theme to the dot.rural

[3] http://www.bbc.co.uk/partnersandsuppliers/connectedstudio/

Digital Economy Hub; award reference EP/G066051/1, and the award made by UK Economic & Social Research Council (ESRC); award reference ES/M001628/1).

8. REFERENCES

[1] G. Álvaro and J. Álvaro. A linked data movie quiz: the answers are out there, and so are the questions [blog post]. `http://bit.ly/linkedmovies`, 2010.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[3] J. Chang, S. Gerrish, C. Wang, J. L. Boyd-graber, and D. M. Blei. Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems*, pages 288–296, 2009.

[4] D. Damljanovic, D. Miller, and D. O'Sullivan. Learning from quizzes using intelligent learning companions. In *WWW (Companion Volume)*, pages 435–438, 2013.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009.

[6] C. Fellbaum, editor. *WordNet: an electronic lexical database*. MIT Press, 1998.

[7] T. L. Griffiths, M. Steyvers, and J. B. Tenenbaum. Topics in semantic representation. *Psychological review*, 114(2):211, 2007.

[8] J. Hees, T. Roth-Berghofer, R. Biedert, B. Adrian, and A. Dengel. Betterrelations: Collecting association strengths for linked data triples with a game. In *Search Computing*, volume 7538 of *LNCS*, pages 223–239. 2012.

[9] A. Huang. Similarity measures for text document clustering. In *Proceedings of the Sixth New Zealand Computer Science Research Student Conference (NZCSRSC2008), Christchurch, New Zealand*, pages 49–56, 2008.

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

[11] C. Leacock and M. Chodorow. Combining local context and wordnet similarity for word sense identification. *WordNet: An electronic lexical database*, 49(2):265–283, 1998.

[12] C. Lin, Y. He, C. Pedrinaci, and J. Domingue. Feature lda: a supervised topic model for automatic detection of web api documentations from the web. In *The Semantic Web–ISWC 2012*, pages 328–343, 2012.

[13] A. Passant. Measuring semantic distance on linking data and using it for resources recommendations. In *AAAI Spring Symposium: Linked Data Meets Artificial Intelligence*, 2010.

[14] G. Á. Rey, I. Celino, P. Alexopoulos, D. Damljanovic, M. Damova, N. Li, and V. Devedzic. Semi-automatic generation of quizzes and learning artifacts from linked data. In *Linked Learning 2012: 2nd International Workshop on Learning and Education with the Web of Data, at the World Wide Web Conference 2012 (WWW2012)*, 2012.

[15] K. Siorpaes and M. Hepp. Games with a purpose for the semantic web. *IEEE Intelligent Systems*, 23(3):50–60, 2008.

[16] J. L. Szalma, J. S. Warm, G. Matthews, W. N. Dember, E. M. Weiler, A. Meier, and F. T. Eggemeier. Effects of sensory modality and task duration on performance, workload, and stress in sustained attention. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 46(2):219–233, 2004.

[17] P. D. Turney. Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings of the 12th European Conference on Machine Learning (ECML)*, pages 491–502, London,UK, 2001.

[18] J. Waitelonis, N. Ludwig, M. Knuth, and H. Sack. WhoKnows? - evaluating linked data heuristics with a quiz that cleans up dbpedia. *International Journal of Interactive Technology and Smart Education (ITSE)*, 8:236–248, 2011.

[19] L. Wolf, M. Knuth, J. Osterhoff, and H. Sack. Risq! renowned individuals semantic quiz: A jeopardy like quiz game for ranking facts. In *Proceedings of the 7th International Conference on Semantic Systems*, I-Semantics '11, pages 71–78, 2011.

[20] Z. Wu and M. Palmer. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics (ACL)*, pages 133–138, 1994.