Is Brad Pitt Related to Backstreet Boys? Exploring Related Entities

Nitish Aggarwal, Kartik Asooja, Paul Buitelaar, and Gabriela Vulcu

Unit for Natural Language Processing
Insight-centre, National University of Ireland
Galway, Ireland
firstname.lastname@insight-centre.org

Abstract. More than 40% web search queries pivot around a single entity, which can be interpreted with the available knowledge bases such as DBpedia and Freebase. With the advent of these knowledge repositories, major search engines are interested in suggesting the related entities to the users. These related entities provide an opportunity to the users to extend their knowledge. Related entities can be obtained from knowledge bases by retrieving the directly linked entities. However, many popular entities may have more than 1,000 directly connected entities without defining any associativity weight, and these knowledge bases mainly cover some specific types of relations. For instance, "Brad Pitt" can be considered related to "Tom Cruise" but they are not directly connected in DBpedia graph with any relation. Therefore, it is worth finding the related entities beyond the relations explicitly defined in knowledge graphs, and a ranking method is also required to quantify the relations between the entities. In this work, we present Entity Relatedness Graph (EnRG) that provides a ranked list of related entities to a given entity, and clusters them by their DBpedia types. Moreover, EnRG helps a user in discovering the provenance of implicit relations between two entities. With the help of EnRG, one can easily explore even the weak relations that are not explicitly included in the available knowledge bases, for instance, Why is "Brad Pitt" related to "Backstreet Boys"?

1 Introduction

Since major web search engines are interested in increasing the users' engagement by giving an opportunity for the users to explore information about entities that are related to their initial intent. For instance, Google and Yahoo! recommend the related persons, locations, organizations, movies, songs and events, by using their knowledge graphs. With the help of such recommendations, users can easily explore and expand their knowledge by starting with their initial interest. The available knowledge graphs such as DBpedia and Freebase consist of several entities that are explicitly connected with a given entity through different relations. However, entities may easily connect to many different entities in the knowledge bases without any quantification of the connection strength resulting into a huge number of connections which are hard to track and traverse. Moreover, these knowledge bases cover only limited types of relations. It could be easily possible that even two popular entities which are not directly connected in DBpedia graph, exhibit some relation/connection mentioned in the textual content of Wikipedia articles. Therefore, one should be able to find the related entities beyond the explicitly defined relations in knowledge graphs. It is also required to rank these relations in order to select the top related entities for a better retrieval and recommendation.

This work presents an entity relatedness graph (EnRG)[1] that provides a ranked list of related entities to a given entity. These related entities help users in extending their knowledge around their initial interest. EnRG considers every Wikipedia article topic as an entity. Similar to DBpedia, every node in EnRG represents an entity described in a Wikipedia article. The edges between the nodes reflect the relatedness score between the corresponding entities in EnRG. Wikipedia contains more than 4.1 million entities. Therefore, to build this graph, we need to compute the relatedness scores between 16.8 trillions (4.1 million x 4.1 million) of entity-pairs. Every Wikipedia article has a corresponding DBpedia page that provides further exploration for different relations of the entity. DBpedia defines rdf:type of every Wikipedia entity, which allows us to get the ranked list for every type. The rdf:type information enable us to group the related entities under person, company, location, movie and others types. For instance, if we are interested to find the related entities of "Brad Pitt", we would obtain "Angelina Jolie", "George Clooney" and "Jennifer Aniston" as top related persons; "World War Z" and "The Curious Case of Benjamin" as top related movies; and other types of entities such as television show "Friends" and the book "Moneyball", and many more. We evaluate EnRG on a gold standard dataset containing a ranked list of 20 related entities for 20 different entity queries.

Major search engines recommend the related entities utilizing various resources such as query logs, query session, user click-logs, and knowledge graph [2, 11], however their datasets are not publicly available and they do not provide the related entities belonging to weakly related types. On the contrary, EnRG uses publicly available resources like Wikipedia and its derived knowledge base DBpedia, and EnRG provides many different types of related entities. As an example, the search engines recommend mostly the related persons and movies for the query "Brad Pitt", as "Brad Pitt" is popular as a film actor. But, an entity can be related to other entities through other roles than the most popular ones. A significant number of entities related to "Brad Pitt" may not necessarily fall under person or movie types. These related entities provide an opportunity for the users to explore more about their interests through implicit and non-popular relationships between the entities. It may not be very obvious for everyone that the music band "Backstreet boys" is related to "Brad Pitt", and the users could be interested to know the reasons for any connection provided by EnRG. Therefore,

[1] EnRG Link: http://server1.nlp.insight-centre.org:8080/enrg/

EnRG provides the users an opportunity to explore the provenance information about the retrieved related entities. It shows that A.J. McLean[2] from "Backstreet Boys" played a similar character as of "Brad Pitt" in his version of "Fight Club".

2 Related Work

EnRG can be seen as a search assistance system that retrieves several ranked lists of related entities classified under different types, and provides a further exploration on the results to the search query. Classical approaches [6] for search assistance provides suggestions for related queries to a given query by using co-occurrence statistics from query logs and query sessions data. However, recent work [8] has identified that more than 40% search queries revolve around a single entity, which encourages us to develop an interactive and intelligent search around entities. Moreover, it gives an opportunity for the users to explore more by recommending related knowledge to users' initial interest. Recently, entity recommendation has received much attention in web search. Major search engines recently published their work on recommending related entities to the user search query. Blanco et al. [2] introduced Spark that links a user search query to an entity in knowledge base and provides a ranked list of related entities to further exploration. Spark uses different features from several datasets such as Flickr, Yahoo! query logs and Yahoo Knowledge graph. Spark tunes the parameters by using learning to rank. Similarly, Yu et al. [11] proposed a personalized entity recommendation which uses several features extracted from user click logs provided through Bing search. Search engines specific datasets like query logs and user click logs are not publicly available, and generating a training data for such solutions is expensive and not trivial.

Another important aspect of the entity recommendation is the relatedness measure for quantifying the relationship. For instance, in order to decide if "Brad Pitt" is more related to "Angelina Jolie" than "Tom Cruise", it requires some relatedness measure which can measure this on the basis of their connections in the knowledge graphs or co-occurrence in the textual documents etc. Wikipedia and its derived knowledge bases like DBpedia, YAGO and FreeBase provide immense amount of information about millions of entities. The advent of this knowledge about persons, locations, products, events etc. introduces numerous opportunities to develop entity relatedness measures. Strube and Ponzetto [9] proposed WikiRelate that exploits the Wikipedia link structure to compute the relatedness between Wikipedia concepts. WikiRelate counts the edges between two concepts and considers the depth of a concept in the Wikipedia category structure. Ponzetto and Strube [7] adapted the WordNet-based measures to Wikipedia for obtaining the advantages of its constantly growing vocabulary. Witten and Milne [10] applied Google distance metric [3] on incoming links to Wikipedia. These approaches perform only for the entities which appear on

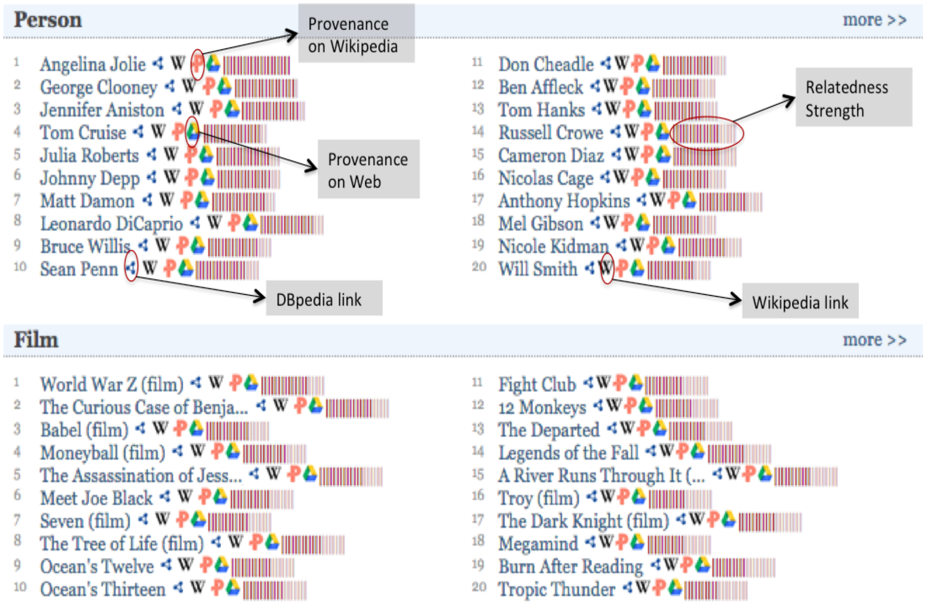
[2] `http://en.wikipedia.org/wiki/A._J._McLean`

Wikipedia. KORE [5] eliminates this issue by computing the relatedness scores between the context of two entities. It observes the partial overlaps between the concepts (key-phrases) appearing in the contexts of the given two entities. KORE improves over other existing algorithms. However, it takes only the surface forms of the concepts appearing in the context and does not utilize their background knowledge.

3 Entity Relatedness Graph (EnRG)

Entity Relatedness Graph (EnRG) is constructed by calculating the entity relatedness scores between every entity pair. We calculate entity relatedness score by using our recent work Wikipedia-based Distributional Semantics for Entity Relatedness (DiSER) [1]. DiSER builds the semantic profile of an entity by using the high dimensional concept space derived from Wikipedia articles. DiSER generates a high dimensional vector by taking every Wikipedia article as a vector dimension, and the associativity weight of an entity with the concept as the magnitude of the corresponding dimension. To measure the semantic relatedness between two entities, it simply calculates the cosine [12] score between their corresponding DiSER vectors. It retrieves a list of the relevant Wikipedia concepts and rank them according to their relevance scores with the given entity. DiSER considers only hyperlinks in Wikipedia, thus keeping all the canonical entities that appear with hyperlinks in Wikipedia articles. The tf-idf weight of an entity with every Wikipedia article is calculated and used to build a semantic profile. The semantic profile of an entity is represented by the retrieved Wikipedia concepts sorted by their tf-idf scores. For instance, there is an entity e, DiSER builds a semantic vector v, where $v = \sum_{i=0}^{N} a_i * c_i$ and c_i is i^{th} concept in the Wikipedia concept space, and a_i is the tf-idf weight of the entity e with the concept c_i. Here, N represents the total number of Wikipedia concepts.

Wikipedia contains more than 4.1 million entities. Therefore, to build EnRG graph, we calculate the relatedness scores between 16.8 trillions (4.1 million x 4.1 million) of entity-pairs. DiSER builds the distributional vector over Wikipedia articles, therefore, every DiSER vector has 4.1 million dimensions. It can be seen as a sparse square matrix of order 4.1 million. To reduce the number of computations, we keep only the top 1000 dimensions and converge the remaining dimensions' score to zero. We need to calculate the cosine scores between all the rows. In order to produce 16.8 trillion scores, a very fast and efficient computing is required. Even if the system is able to process more than half a millions entity-pairs per second, the complete process will take more than a year. Therefore, we applied a pruning technique which only calculates the DiSER score if it would be a non-zero value. We collect all the possible related entities with non-zero scores for a given entity. Since we take only top 1,000 articles to build the vector, the entities not appearing in the content of the top 1,000 articles for a given entity would produce a zero relatedness score with that entity. For instance, if DiSER takes only the top 2 articles to calculate the relatedness score, and we want to retrieve all the entities having a non-zero relatedness score with "Apple Inc.", we would obtain all the entities such as "Steve Jobs", "iPad" and "OS X" as they appear in the content of top 2 articles of "Apple Inc." and would not retrieve entities like "Samsung" and "Motorola" as they do not appear in top 2 articles. We obtained around 10K related entities for every individual entity. Therefore, we calculate DiSER scores for only 4.1 billion entity-pairs, and this reduces the comparisons by 99.8%. Our system takes around 48 hours to build the EnRG graph with 25K comparisons per second. Similar to the major search engines, EnRG graph provides a rank list of related entities for a given entity. Moreover, it also categorizes the related entities in many different classes using their DBpedia types. Therefore, a user can obtain a rank list of different type of related entities such as "Person", "Film", "Company" and others. Figure 1 shows a snapshot of our EnRG interface, which illustrates a rank list of related people and films to Brad Pitt.



Person more >>

1 Angelina Jolie
2 George Clooney
3 Jennifer Aniston
4 Tom Cruise
5 Julia Roberts
6 Johnny Depp
7 Matt Damon
8 Leonardo DiCaprio
9 Bruce Willis
10 Sean Penn
11 Don Cheadle
12 Ben Affleck
13 Tom Hanks
14 Russell Crowe
15 Cameron Diaz
16 Nicolas Cage
17 Anthony Hopkins
18 Mel Gibson
19 Nicole Kidman
20 Will Smith

Film more >>

1 World War Z (film)
2 The Curious Case of Benja...
3 Babel (film)
4 Moneyball (film)
5 The Assassination of Jess...
6 Meet Joe Black
7 Seven (film)
8 The Tree of Life (film)
9 Ocean's Twelve
10 Ocean's Thirteen
11 Fight Club
12 12 Monkeys
13 The Departed
14 Legends of the Fall
15 A River Runs Through It (...
16 Troy (film)
17 The Dark Knight (film)
18 Megamind
19 Burn After Reading
20 Tropic Thunder

Fig. 1. Rank list of related people and films to Brad Pitt

4 Evaluation

We implemented EnRG by using the snapshot of English Wikipedia from 1st Aug, 2014. This snapshot consists of 13,872,614 articles, in which 5,659,383 are Wikipedia redirects. We obtained 4,102,442 articles after removing all the

Wikipedia redirects, namespaces and short articles. In order to compare the quality of relatedness scores obtained from our approach, we performed our experiments on the same gold standard dataset KORE [5] that has been used by state of the art methods for calculating entity relatedness. The KORE dataset consists of 21 queries which correspond to the entities from the YAGO knowledge base. Every query has a ranked list of 20 related entities. These queries are selected from 4 different domains: IT companies, Hollywood celebrities, video games, and television series. The KORE dataset consists of 420 entity pairs and their relative relatedness scores.

EnRG graph is built by using our recent work DiSER on entity related measures. To determine the performance of DiSER model, we performed experiments with a similar model i.e. ESA [4], which calculates the relatedness score by using all the textual content in Wikipedia articles rather than considering only the hyperlinks. Other existing methods [5, 10] of calculating entity relatedness utilize content associated with the given entity. The best performing state of the art method KORE [5] calculates entity relatedness by computing the overlap between important key-phrases which appear in the corresponding Wikipedia articles. However, KORE does not make use of distributional or co-occurrence information about an entity. We calculated relatedness scores for all entity pairs provided by the KORE gold standard. Since the gold standard dataset consists of human judgement about the ranking of 20 entities for each entity query, the previous methods reported their accuracy by using Spearman Rank correlation. Therefore, we calculated Spearman Rank correlation between the gold standard dataset and the results obtained from our experiments. Table 1 shows the results. DiSER is compared with three state of the art methods: WLM [10], KORE [5] and ESA [4]. DiSER improves over ESA by 20% which shows the importance of considering only hyperlinks in building distributional representation of an entity. DiSER outperforms all the state of the art methods.

Entity Relatedness Measures	**Spearman Rank Correlation with human**
WLM	0.610
KORE	0.698
ESA	0.691
DiSER	**0.781**

Table 1. Spearman rank correlation of relatedness measures with gold standard

5 Conclusion and Future Work

We presented EnRG, which is a big graph of connected entities based on Wikipedia. Given an entity, it retrieves ranked lists of related entities for different types defined in DBpedia like person, movie, companies, events etc. It provides a further exploration on the results resulting into a a kind of entity exploration system. It also provides the provenance information for any connection between the entities. As a future step, we plan to extend it as a multilingual EnRG system which uses the Wikipedia in other languages. We would also provide it as a REST service allowing the developers and researchers to use it in their applications or research.

References

1. N. Aggarwal and P. Buitelaar. Wikipedia-based distributional semantics for entity relatedness. In *2014 AAAI Fall Symposium Series*, 2014.
2. R. Blanco, B. B. Cambazoglu, P. Mika, and N. Torzec. Entity recommendations in web search. In *International Semantic Web Conference (2)*, pages 33–48, 2013.
3. R. L. Cilibrasi and P. M. Vitanyi. The google similarity distance. *Knowledge and Data Engineering, IEEE Transactions on*, 19(3):370–383, 2007.
4. E. Gabrilovich and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th international joint conference on Artifical intelligence*, IJCAI'07, pages 1606–1611, San Francisco, CA, USA, 2007. Morgan Kaufmann Publishers Inc.
5. J. Hoffart, S. Seufert, D. B. Nguyen, M. Theobald, and G. Weikum. Kore: Keyphrase overlap relatedness for entity disambiguation. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 545–554. ACM, 2012.
6. R. Jones, B. Rey, O. Madani, and W. Greiner. Generating query substitutions. In *Proceedings of the 15th international conference on World Wide Web*, pages 387–396. ACM, 2006.
7. S. P. Ponzetto and M. Strube. Knowledge derived from wikipedia for computing semantic relatedness. *J. Artif. Intell. Res.(JAIR)*, 30:181–212, 2007.
8. J. Pound, P. Mika, and H. Zaragoza. Ad-hoc object retrieval in the web of data. In *Proceedings of the 19th international conference on World wide web*, pages 771–780. ACM, 2010.
9. M. Strube and S. P. Ponzetto. Wikirelate! computing semantic relatedness using wikipedia. In *AAAI*, volume 6, pages 1419–1424, 2006.
10. I. Witten and D. Milne. An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In *Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy, AAAI Press, Chicago, USA*, pages 25–30, 2008.
11. X. Yu, H. Ma, B.-J. P. Hsu, and J. Han. On building entity recommender systems using user click log and freebase knowledge. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, WSDM '14, pages 263–272, New York, NY, USA, 2014. ACM.
12. J. Zobel and A. Moffat. Exploring the similarity space. In *ACM SIGIR Forum*, volume 32, pages 18–34. ACM, 1998.

6 Appendix

This section describes how we fulfill the requirements of **Open Track Submission**.

6.1 Minimal Requirements

End-User Application: EnRG is an entity explorer system containing an intuitive user interface that provides the users an opportunity to explore about related persons, movies, TV shows and many more. DBpedia is used to define the different entity types in EnRG. Researchers and developers can also utilize EnRG in their applications by querying it with the help of a REST API [3] built over EnRG, which provides the results in a standard Json format.

Diverse, Heterogeneous, Real world data: EnRG essentially integrates the entity relatedness information learned using Wikipedia corpus co-occurrence information about the entities, with the structural information embedded in DBpedia or knowledge bases. EnRG categorizes the different entities on the basis of DBpedia ontology. DBpedia provides further specific relations between the related entities from EnRG. Wikipedia contains more than 4.1 million entities, and EnRG contains the relatedness scores between every entity pair on English Wikipedia/DBpedia. This shows the potential of the application in real world.

Semantic information: Given an entity, this application provides the related and ranked entities classified into their types. It also provides the provenance information for the relationships. By using EnRG, users can extend their knowledge about the related entities and their initial interest. With the information about entity types and DBpedia entity links, users also have the opportunity to explore further knowledge for specific class type and from DBpedia.

6.2 Additional Desirable Features

The application provides an attractive and functional web interface. The interface provides an interactive user experience for further exploration about different entities. It provides a visualization for the provenance information about the relation between two entities. EnRG also provides a REST API for interacting with EnRG service.

EnRG is built over all the Wikipedia entities, and the graph is updated every month. The application is scalable as it take less than 48 hours to build this big graph.

We performed a rigorous evaluation of our system. EnRG is evaluated on standard benchmark datasets and it is shown outperforming the state of the art methods. EnRG outperforms other existing methods and achieved a high accuracy. It makes use of contextual information to interpret the relatedness strength. EnRG can also provide related entities for a topic like “Semantic Web” and “Natural Language Processing”.

[3] EnRG Link: http://server1.nlp.insight-centre.org:8080/enrg/