Who Are the American Vegans Related to Brad Pitt? Exploring Related Entities

Nitish Aggarwal
Insight-Centre
National University of Ireland
Galway, Ireland
firstname.lastname@insight-centre.org

Kartik Asooja
Insight-Centre
National University of Ireland
Galway, Ireland
firstname.lastname@insight-centre.org

Housam Ziad
Insight-Centre
National University of Ireland
Galway, Ireland
firstname.lastname@insight-centre.org

Paul Buitelaar
Insight-Centre
National University of Ireland
Galway, Ireland
firstname.lastname@insight-centre.org

ABSTRACT

In this demo, we present Entity Relatedness Graph (EnRG), a focused related entities explorer, which provides the users with a dynamic set of filters and facets. It gives a ranked lists of related entities to a given entity, and clusters them using the different filters. For instance, using EnRG, one can easily find the American vegans related to Brad Pitt or Irish universities related to Semantic Web. Moreover, EnRG helps a user in discovering the provenance for implicit relations between two entities. EnRG uses distributional semantics to obtain the relatedness scores between two entities.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Search Process, Information Filtering

Keywords

Entity Relatedness, Entity Recommendation, Entity Graph, Distributional Semantics, Wikipedia

1. INTRODUCTION

Today, major search engines suggest related entities to the users' queries, which provides the users an opportunity to explore more and extend their knowledge. Recent statistics on search queries suggest that more than 40% of the search queries revolve around a single entity, which encourages us to develop an interactive and intelligent search around entities [8]. Google and Yahoo! recommend the related persons, locations, organizations, movies, songs and events by using their knowledge graphs [3]. The publicly available structured knowledge resources such as DBpedia and Freebase consist of limited types of relations which can be defined between two entities, consequently missing many connections which might appear in the real world. Many such missing connections can be explored through unstructured knowledge sources like news articles or Wikipedia articles. Moreover, the structured knowledge resources do not provide any quantification of the relationship strength, which makes it hard for humans to easily explore and traverse the relationships.

Therefore, in this work, we present Entity Relatedness Graph (EnRG)[1], which provides the users an easy exploration over the related entities beyond the explicitly defined relations in knowledge graphs. It provides a ranking of the connections between the entities in order to select the top related entities for a better retrieval and recommendation. The types of entities suggested by the current search engines are generally abstract and limited to People, Movie, and others. On the contrary, EnRG allows an extensive retrieval on the basis of dynamic facets and filters using DBpedia. EnRG considers every Wikipedia article topic as an entity except the Wikipedia pages referring to lists, disambiguation pages, and redirects. Similar to DBpedia, every node in EnRG represents a Wikipedia based article. The edges between the nodes reflect the relatedness scores between the corresponding entities in EnRG.

EnRG can also 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.

2. RELATED WORK

Classical approaches [6] for search assistance provide suggestions for related queries to a given query by using co-occurrence statistics from query logs and query session data. Recently, entity recommendation has received much attention in web search. Major search engines have recently published their work on recommending related entities to the user search query [3] [12]. In addition to the knowledge graph, search engines also utilize various resources such as

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

WWW 2015 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2742851.

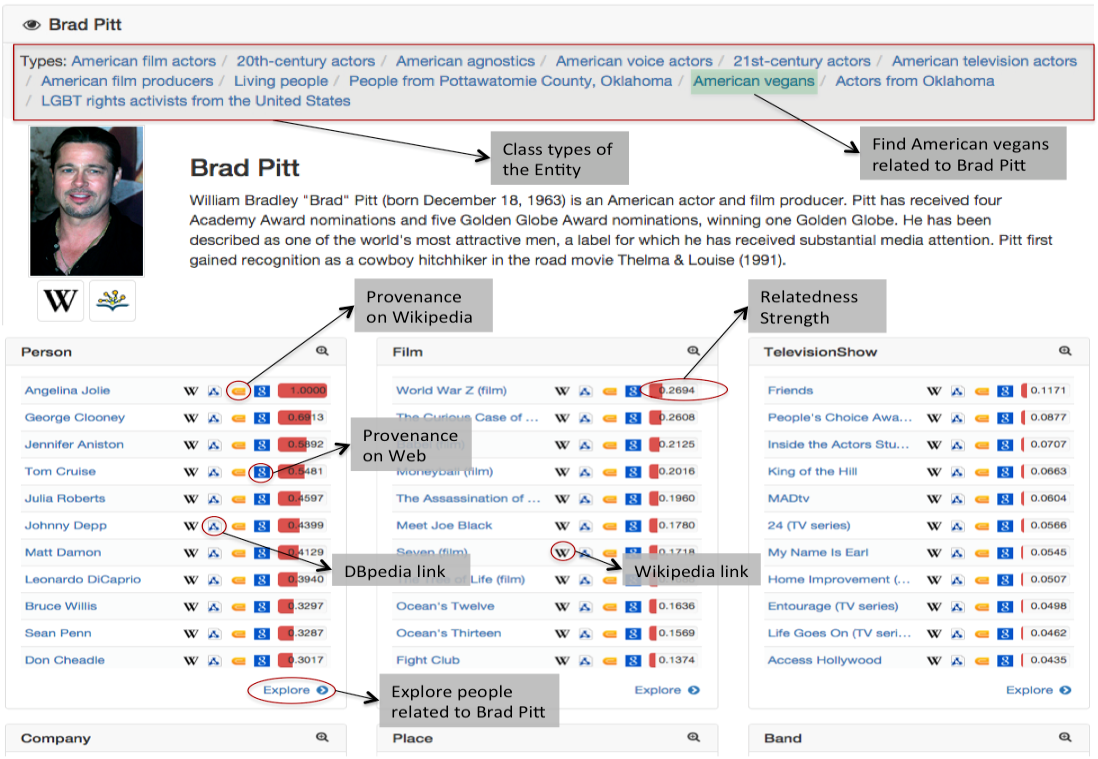

Figure 1: Ranked entities to Brad Pitt

query logs, query session, and user click-logs for recommending the related entities. Blanco et al. [3] introduced Spark that links a user search query to an entity in knowledge base and then provides a ranked list of related entities for further exploration. Spark uses different features from several datasets such as Flickr, Yahoo! query logs and Yahoo Knowledge graph. It tunes the parameters by using learning to rank. Similarly, Yu et al. [12] 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. Unlike such approaches, EnRG uses publicly available resources like Wikipedia and its derived knowledge base DBpedia.

Another important aspect of the entity recommendation is the relatedness measure for quantifying the relationship on the basis of the connections in the knowledge graphs or co-occurrence in the textual documents. 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 [11] applied Google distance metric [4] on incoming links to Wikipedia. These approaches perform only for the entities which appear on Wikipedia. KORE [5] eliminates this issue by computing the relatedness scores between the contexts of two entities. EnRG is based upon Wikipedia-based Distributional Semantics for Entity Relatedness (DiSER) [2], which has been shown to outperform state of the art algorithms for entity relatedness.

3. SYSTEM IMPLEMENTATION

Entity Relatedness Graph (EnRG) is constructed by calculating the entity relatedness scores between every entity pair following our recent work DiSER [2]. In this section, we first present DiSER, and then discuss the implementation of EnRG.

3.1 DiSER

DiSER builds the semantic profile of an entity by using the high dimensional concept space derived from Wikipedia articles. It generates a high dimensional vector by taking every Wikipedia article topic as a vector dimension, and associativity weight of an entity with the topic as the magnitude of the corresponding dimension [2] [1]. To measure the semantic relatedness between two entities, it computes the cosine score between their corresponding DiSER vectors. DiSER considers only the hyperlinks in Wikipedia, thus keeping all the canonical entities that appear with hyperlinks in Wikipedia articles. 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.

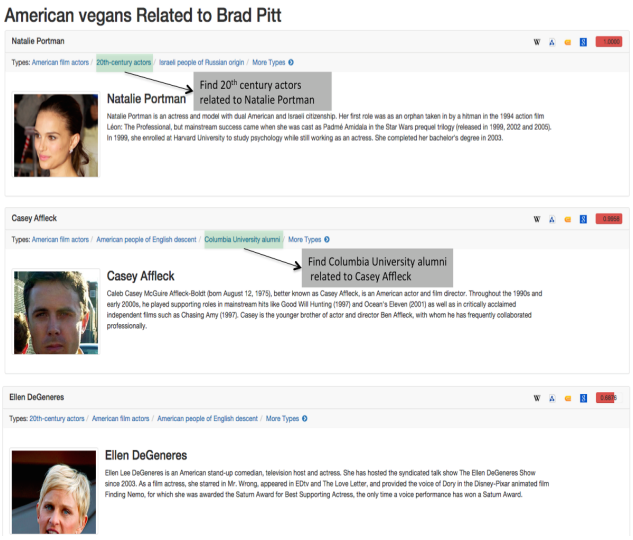

Figure 2: Ranked list of American vegans related to Brad Pitt

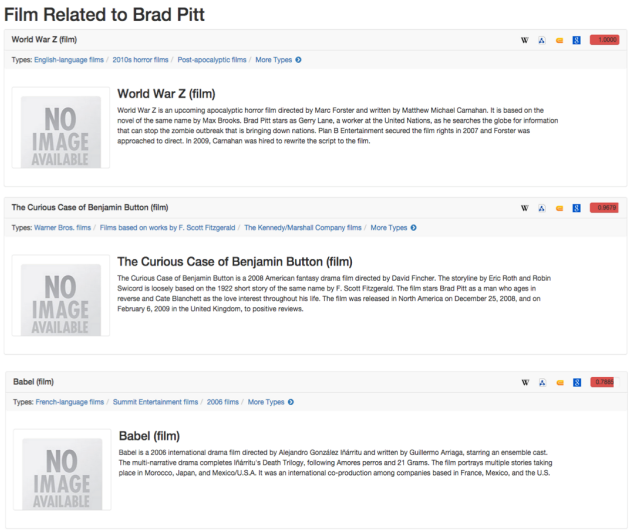

Figure 3: Ranked list of Films related to Brad Pitt

3.2 EnRG

Wikipedia[2] 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. 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 in the DiSER vector according to their associativity scores with the entity, which converges the remaining dimensions' scores to zero.

[2]It is a snapshot of English Wikipedia from October 2013

In order to produce 16.8 trillion scores, a very fast and efficient computing is required. 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 the 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 the entities such as "Steve Jobs", "iPad" and "OS X" as they appear in the content of the top 2 articles of "Apple Inc.", while we would miss entities like "Samsung" and "Motorola" as they do not appear in the top 2 articles. We obtain around 10K related entities for every individual entity in EnRG. 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 seconds.

4. DEMO

This demonstration based on EnRG provides the users with a dynamic set of filters and facets for exploring the related entities with the help of DBpedia. Every Wikipedia article has a corresponding DBpedia page that provides further exploration for different relations of the entity. DBpedia defines rdf:type[3] of every Wikipedia entity, which allows us to get the ranked list for each type. These types include classes mainly from DBpedia ontology[4] and YAGO [10]. DBpedia ontology covers abstract types like Person, Company, Location, Movie and others, while YAGO also provides very specific types like American film actors, People from Manhattan, etc. This information enable us to group the related entities under different types. Figure 1 shows a snapshot

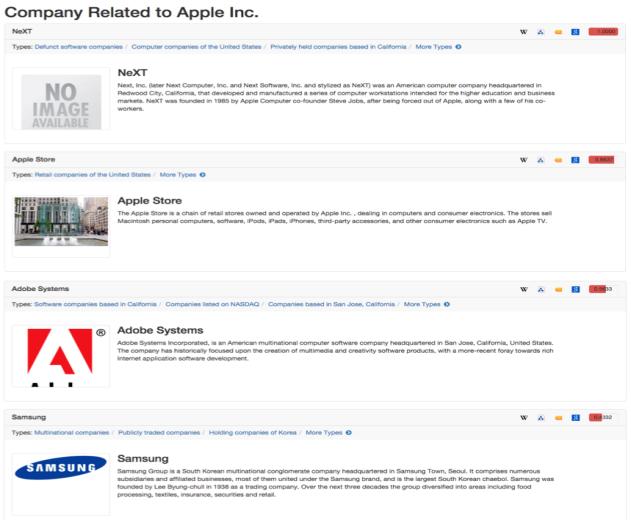

Figure 4: Ranked list of Companies related to Apple Inc.

of the EnRG interface, which illustrates ranked lists of re-

[3]http://www.w3.org/1999/02/22-rdf-syntax-ns#type
[4]http://wiki.dbpedia.org/Ontology

lated entities to Brad Pitt, categorized under different abstract types. Just below the main heading "Brad Pitt", it shows the different specific classes such as American vegans, this entity can belong to. The figure also shows different functionalities of the demonstration. On clicking any entity type, it would list all the entities listed under that type, for instance, using such filters for Brad Pitt, one can easily navigate to find the American vegans related to Brad Pitt as shown in Figure 2. To explore the related films to Brad Pitt, one can simply click the "Explore" link at the bottom of the list of related films, which is shown in Figure 3. The application also presents two types of provenance information, one based on Wikipedia giving the Wikipedia articles which contain both the related entities, while the other performs a Google web search giving the provenance information from the web. It also shows the relatedness strength quantifying the entity relatedness. The application also gives query suggestions referring to related entities based on the searched query string, for instance, if someone searches just using the string "Apple", then it shows the related entities for Apple fruit, but also gives other query suggestions like Apple Inc. and Apple Records. Selecting the query suggestion "Apple Inc." would lead to the related entities of Apple Inc., through which we can easily navigate to the related companies as shown in Figure 4.

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 DBpedia types. It provides further exploration on the results based on the specific types given by YAGO classes. As a future step, we plan to extend it as a multilingual EnRG system which uses the Wikipedia in other languages. We also provide it as a REST service allowing the developers and researchers to use it in their applications or research.

6. ACKNOWLEDGEMENTS

This work has been funded by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI /12/RC/2289 (INSIGHT).

7. REFERENCES

[1] N. Aggarwal, K. Asooja, P. Buitelaar, and G. Vulcu. Is brad pitt related to backstreet boys? exploring related entities. In *Semantic Web Challenge ISWC*, 2014.

[2] N. Aggarwal and P. Buitelaar. Wikipedia-based distributional semantics for entity relatedness. In *2014 AAAI Fall Symposium Series*, 2014.

[3] 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.

[4] R. L. Cilibrasi and P. M. Vitanyi. The google similarity distance. *Knowledge and Data Engineering, IEEE Transactions on*, 19(3):370–383, 2007.

[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] F. M. Suchanek, G. Kasneci, and G. Weikum. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, pages 697–706. ACM, 2007.

[11] 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.

[12] 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.