Result Diversity for RDF Search

Hiba Arnaout and Shady Elbassuoni
Computer Science Department, American University of Beirut, Bliss Street, Beirut, Lebanon

Keywords: Diversity, Novelty, RDF-graphs, Evaluation.

Abstract: RDF repositories are typically searched using triple-pattern queries. Often, triple-pattern queries will return too many results, making it difficult for users to find the most relevant ones. To remedy this, some recent works have proposed relevance-based ranking-models for triple-pattern queries. However it is often the case that the top-ranked results are homogeneous. In this paper, we propose a framework to diversify the results of triple-pattern queries over RDF datasets. We first define different notions for result diversity in the setting of RDF. We then develop an approach for result diversity based on the Maximal Marginal Relevance. Finally, we develop a diversity-aware evaluation metric based on the Discounted Cumulative Gain and use it on a benchmark of 100 queries over DBPedia.

1 INTRODUCTION

The continuous growth of knowledge-sharing communities like Wikipedia and the recent advances in information extraction have made it possible to build large knowledge-bases such as YAGO (Suchanek et al., 2008) and DBPedia (Auer et al., 2007). These knowledge bases consist of billions of facts represented in the W3C semantic model RDF (RDF, 2004). Querying these RDF knowledge bases or repositories is typically done by triple-pattern-based query languages such as SPARQL (SPARQL, 2008).

For example, consider the triple-pattern query: *"?m director ?d; ?m genre Comedy"*. This query consists of two triple patterns and aims to find directors of comedy films. For example, running our example query over DBPedia returns 1073 results. Table 1 shows 5 examples of these 1073 results.

Users usually prefer seeing a ranked result-list rather than a list of unranked matches (Chaudhuri et al., 2006). Recently, some approaches have been proposed to rank the results of triple-pattern queries. For example, the work in (Elbassuoni et al., 2009) proposed a ranking approach based on language models to rank query results. In (Dali et al., 2012), the authors proposed a learning-to-rank approach that uses query-independent features. Another example is the work in (Kasneci et al., 2008) where the authors use several notions such as confidence, informativeness and compactness to rank query results. Table 2 shows the top-5 *ranked* results for our example query when run over DBPedia using the ranking model of (Elbassuoni et al., 2009). As shown in Table 2, the top ranked results are all about famous movies and directors, as compared to those in Table 1 which are about unpopular movies and directors.

Table 1: 5 example results from DBPedia for the query *"?m director ?d; ?m genre Comedy"*.

Subject	Predicate	Object
Model_Ball	director	Scott_Zarakin
Model_Ball	genre	Comedy
All_the_Wrong_Places	director	Martin_Edwards
All_the_Wrong_Places	genre	Comedy
Friends_(2002_film)	director	M._D._Sridhar
Friends_(2002_film)	genre	Comedy
Double_or_Nothing	director	Roy_Mack
Double_or_Nothing	genre	Comedy
President's_Day	director	Chris_LaMartina
President's_Day	genre	Comedy

Table 2: Top-5 ranked results from DBPedia for the query *"?m director ?d; ?m genre Comedy"*.

Subject	Predicate	Object
Annie_Hall	director	Woody_Allen
Annie_Hall	genre	Comedy
Dumb_and_Dumber	director	Perer_Farrelly
Dumb_and_Dumber	genre	Comedy
Sleeper	director	Woody_Allen
Sleeper	genre	Comedy
Husbands_and_Wives	director	Woody_Allen
Husbands_and_Wives	genre	Comedy
Muppets_Most_Wanted	director	James_Bobin
Muppets_Most_Wanted	genre	Comedy

Arnaout, H. and Elbassuoni, S.
Result Diversity for RDF Search.
DOI: 10.5220/0006046402490256
In *Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2016) - Volume 1: KDIR*, pages 249-256
ISBN: 978-989-758-203-5

While result ranking goes a long way in improving the user satisfaction, it is often the case that the top-ranked results are dominated by one aspect of the query. This is a common problem in IR in general (Carbonell and Goldstein, 1998). For example, consider our example query. As can be seen from Table 2, a large number of the top-5 results are movies by the same director, namely Woody Allen.

To achieve the tradeoff between the relevance and diversity of a result, we rely on the Maximal Marginal Relevance (MMR) approach (Carbonell and Goldstein, 1998) which we adapt to the RDF setting.

To evaluate the effectiveness of our diversity approach and to compare different notions of diversity, we define a new evaluation metric based on the Discounted Cumulative Gain (DCG) (Jrvelin and Keklinen, 2002).

Our contributions can be summarized as follows:

- We provide the first formal definition of result diversity in the context of RDF Search.
- We develop the first diversity-aware ranking model for RDF Search.
- We design a new diversity-aware evaluation metric for RDF search.
- We built the first evaluation benchmark on DBPedia that can be used to evaluate ranking models and result diversity approaches for RDF search.

2 RELATED WORK

Result diversity for document retrieval has gained much attention in recent years. The work in this area deals primarily with unstructured and semi-structured data (Agrawal et al., 2009; Carbonell and Goldstein, 1998; Chen and Karger, 2006; Clarke et al., 2008; Gollapudi and Sharma, 2009; Zhai et al., 2003). Most of the techniques perform diversification by optimizing a bi-criteria objective function that takes into consideration both result relevance as well as result novelty with respect to other results. Gollapudi and Sharma (Gollapudi and Sharma, 2009) presented an axiomatic framework for this problem and studied various objective functions that can be used to define such optimization problem. They proved that in most cases, such problem is hard to solve and proposed several approximation algorithms to solve such problem. Carbonell and Goldstein introduced the Maximal Marginal Relevance (MMR) method (Carbonell and Goldstein, 1998) which is one approximation solution to such optimization problem. Zhai et al. (Zhai et al., 2003) studied a similar approach within the framework of language models and derived an MMR-based loss function that can be used to perform diversity-aware ranking. Aragwal et al. (Agrawal et al., 2009) assumed that query results belong to different categories and they proposed an objective function that tries to trade off the relevance of the results with the number of categories covered by the selected results.

In an RDF setting, where results are constructed at query time by joining triples, we do not have an explicit notion of result categories. We thus adopted the Maximal Marginal Relevance approach (Carbonell and Goldstein, 1998) to the setting of RDF data since it directly utilizes the results to perform diversity rather than explicitly taking the categories of the results into consideration.

Apart from document retrieval, there is very little work on result diversity for queries over structured data. In (Chen and Li, 2007) the authors propose to navigate SQL results through categorization, which takes into account user preferences. In (Vee et al., 2008), the authors introduce a pre-indexing approach for efficient diversification of query results on relational databases. However, they do not take into consideration the relevance of the results to the query.

3 RDF SEARCH

In this paper, we tackle the problem of result diversity for RDF search. To perform an RDF search, we rely on the RDF Xpress search engine (Elbassuoni et al., 2009). RDF Xpress supports three different modes of search, namely purely-structured triple-pattern queries, keyword-augmented triple-pattern queries and automatic query relaxation. For purely-structured queries, the search engine takes as input a triple-pattern query and returns a ranked list of RDF subgraphs matching the given query. The results are ranked based on a language-modeling approach specially developed for the setting of RDF search. For example, Table 2 shows the top-5 results for the query: *"?m director ?d; ?m genre Comedy"*. For keyword-augmented queries, the search engine takes as input a triple-pattern query where one or more triple pattern is augmented with a set of keywords. For example, to find movies directed by Woody Allen that have something to do with New York:*"?m director Woody_Allen [new york]"*. One example result is *"Annie_ Hall director Woody_Allen"*. The story of this film was set in New York city.

Finally, for the case when no results are found for a given triple-pattern query, automatic query relaxation is deployed.

For example, running the query *"?m starring Woody_Allen; ?m musicComposer Dick_Hyman"* over DBPedia will return no results. After combining the results of all relaxed queries using the result-merging approach of RDF Xpress, one sample result is *"Radio_Days writer Woody Allen; Radio_Days musicComposer Dick_Hyman"*.

4 RESULT DIVERSITY

Our main focus in this paper is on result diversity for RDF search. A diversity-aware ranking model should ideally try to produce an ordering or a permutation of the query results such that the top-*k* results are most relevant to the query and at the same time as diverse from each other as possible. This can be cast into an optimization problem where the objective is to produce an ordering that would maximize both the relevance of the top-*k* results and their diversity. The objective function for such an optimization problem is very hard to both quantify and solve and thus most approaches try to solve a simpler closely-related problem known as the top-*k* set selection problem (Gollapudi and Sharma, 2009). The top-*k* set selection problem can be formulated as follows.

Definition 1 (Top-*k* Set Selection)**.** *Let Q be a query and U be its result set. Furthermore, let REL be a function that measures the relevance of a subset of results $S \subseteq U$ with respect to Q and let DIV be a function that measures the diversity of a subset of results $S \subseteq U$. Finally, let f be a function that combines both relevance and diversity. The top-k set selection problem can be solved by finding:*

$$S^* = \underset{S \subseteq U}{\operatorname{argmax}} \; f(Q, S, REL, DIV)$$

such that $|S^*| = k$

The objective function $f(Q,S,REL,DIV)$ is clearly underspecified and in order to the solve this optimization problem, one must clearly specify both the relevance function REL and diversity function DIV, as well as how to combine them. Gollapudi and Sharma (Gollapudi and Sharma, 2009) proposed a set of axioms to guide the choice of the objective function $f(Q,S,REL,DIV)$ and they showed that for most natural choices of the relevance and diversity functions, and the combination strategies between them, the above optimization problem is NP-hard. For instance, one such choice of the objective function is the following:

$$f(Q,S,REL,DIV) = (k-1)\sum_{r \in S} rel(r,Q) + 2\lambda \sum_{r,r' \in S} d(r,r') \quad (1)$$

where $rel(r,Q)$ is a (positive) score that indicates how relevant result r is with respect to query Q (the higher this score is, the more relevant r is to Q) and $d(r,r')$ is a discriminative and *symmetric* distance measure between two results r and r', and λ is a scaling parameter.

The above objective function clearly trades off both relevance of results in the top-*k* set with their diversity (as measured by their average distance). Solving such objective function is again NP-hard, however there exists known approximation algorithms to solve the problem that mostly rely on greedy heuristics (Gollapudi and Sharma, 2009).

In the rest of this section, we follow the same approach to obtain a top-*k* set of relevant and diverse results for queries over RDF knowledge bases. In particular, we optimize the above bi-criteria objective function using a greedy algorithm that uses the Maximal Marginal Relevance (MMR) (Carbonell and Goldstein, 1998) to select the top-*k* set.

4.1 Maximal Marginal Relevance

Carbonell and Goldstein introduced the Maximal Marginal Relevance (MMR) method (Carbonell and Goldstein, 1998) which they use to re-rank a set of pre-retrieved documents U given a query Q.

Definition 2 (Marginal Relevance)**.** *Given a query Q, a set of results U and a subset $S \subset U$, the marginal relevance of a result $r \in U \setminus S$ is equal to: $MR(r,Q,S) = \lambda rel(r,Q) + (1-\lambda)\min_{r' \in S} d(r,r')$ where $rel(r,Q)$ is a measure of how relevant r is to Q, $div(r,r')$ is a symmetric distance measure between r and r' and λ is a weighting parameter.*

The idea behind the marginal relevance metric is very intuitive. Given a query Q and a set of already selected results S, the marginal relevance of a result r is a measure of how much do we gain in terms of both relevance and diversity by adding the result r to the selected set S. To measure how much the result r would contribute to the relevance aspect of S, it is straight forward and we can use the result's relevance to Q. On the other hand, measuring how much result r would contribute to the diversity of S is more involved. The most natural thing to do is to compare r with all the results $r' \in S$ and compute a similarity (or rather dissimilarity) between r and every other result $r' \in S$ and then aggregate these similarities over all the results in S. We do exactly this by assuming there is a distance function that can measure how result r is different from any other result r' and then we use the minimum of the distances of r from all the results $r' \in S$ as a measure of the overall contribution

of result r to the diversity of set S. By maximizing this minimum over a set of results $r \notin S$, we can find the result that when added to S would render it most diverse as compared to any other result.

Given all these considerations, we set the relevance $rel(r,Q)$ to the score of the result r obtained from the RDF Xpress engine . We only assume here that the search engine would rank the results descendingly based on their scores.

We propose next three different notions of diversity and then we explain how we build a subgraph representation that allows us to achieve each such notion.

4.1.1 Resource-based Diversity

In this notion of diversity, the goal is to diversify the different resources (i.e., entities and relations) that appear in the results. This ensures that no one resource will dominate the result set. Recall our example query asking for comedy movies and their directors. Table 2 shows the top-5 subgraphs retrieved for the query using RDF Xpress.

In order to diversify the top-k results of a certain query, we define a language model for each result as follows.

Definition 3 (Resource-based Language Model)**.** *The resource-based language model of result r is a probability distribution over all resources in the knowledge base KB.*

The parameters of the result language model are estimated using a smoothed maximum likelihood estimator as follows:

$$P(w|r) = \alpha\frac{c(w;r)}{|r|} + (1-\alpha)\frac{1}{|Col|} \qquad (2)$$

where w is a resource, Col is the set of all unique resources in the knowledge base, $c(w;r)$ is the number of times resource w occurs in r, $|r|$ is the number of times all resources occur in r, and $|Col|$ is the number of unique resources in the knowledge base. Finally, α is the smoothing parameter.

4.1.2 Term-based Diversity

In this notion of diversity, we are only interested in diversifying the results in terms of the variable bindings. To be able to do this, we define a language model for each result as follows.

Definition 4 (Term-based Language Model)**.** *The term-based language model of result r is a probability distribution over all terms (unigrams) in the knowledge base KB.*

The parameters of the result language model are estimated using a smoothed maximum likelihood estimator as follows:

$$P(w|r) = \alpha\frac{c(w;r)}{|r|} + (1-\alpha)\frac{1}{|Col|} \qquad (3)$$

such that $w \notin QTerms$, where w is a term, $QTerms$ is the list of the terms in the query, Col is the set of all unique terms in the knowledge base, $c(w;r)$ is the number of times term w occurs in r, $|r|$ is the number of times all terms occur in r, and $|Col|$ is the number of unique terms in the knowledge base. Finally, α is the smoothing parameter.

By excluding terms that appear in the original query when representing each result, we ensure that when these representations are later used for diversity, the top-ranked results still stay close to the original user query.

4.1.3 Text-based Diversity

In RDF Xpress, each triple is also associated with a text snippet which can be used to process keyword-augmented queries. A text snippet can be directly utilized to provide diversity among the different results using the MMR measure.

To be able to do this, we define a language model for each result as follows.

Definition 5 (Text-based Language Model)**.** *The text-based language model of a result r is a probability distribution over all the keywords in all the text snippets of all the triples in the knowledge base KB.*

The parameters of the text-based language model is computed using a smoothed maximum-likelihood estimator as follows:

$$P(w|r) = \alpha\frac{c(w;D(r))}{|D(r)|} + (1-\alpha)\frac{1}{|Col|} \qquad (4)$$

where $c(w;D(r))$ is the number of times keyword w occurs in $D(r)$ (the text snippet of subgraph r), $|D(r)|$ is the number of occurrences of all keywords in $D(r)$, and $|Col|$ is the number of unique keywords in the text snippets of all triples in the knowledge base. Finally, α is the smoothing parameter.

4.2 Diversity-aware Re-ranking Algorithm

Finally, we explain how the marginal relevance can be used to provide a diverse-aware ranking of results given a query Q. Let U be the set of ranked results using any regular ranking model (i.e., that depends only on relevance without taking into consideration diversity). The algorithm to re-rank the results works as follows:

Maximal Marginal Relevance Re-ranking Algorithm

1. Initialize the top-k set S with the highest ranked result $r \in U$
2. Iterate over all the results $r \in U \setminus S$, and pick the result r^* with the maximum marginal relevance $MR(r^*, Q, S)$. That is,

$$r^* = \underset{r \in U \setminus S}{\operatorname{argmax}} \; [\lambda rel(r,Q) + (1-\lambda) \min_{r' \in S} d(r,r')] \tag{5}$$

3. Add r^* to S
4. If $|S| = k$ or $S = U$ return S otherwise repeat steps 2, 3 and 4

The distance between two results r and r' is computed as follows:

$$d(r,r') = \sqrt{JS(r||r')}$$

where $JS(r||r')$ is the Jensen-Shannon Divergence (Lin., 1991) between the language models of results r and r' and is computed as follows:

$$\begin{aligned} JS(r||r') &= \frac{KL(r||M) + KL(r'||M)}{2} \\ &= \frac{\sum_i^{terms} r(i) log \frac{r(i)}{M(i)} + r'(i) log \frac{r'(i)}{M(i)}}{2} \end{aligned} \tag{6}$$

where $KL(x||y)$ is the Kullback-Leibler Divergence between two language models x and y, and $M = \frac{1}{2}(r + r')$ is the average of the language models of r and r'.

We opted for using the Jensen-Shannon Divergence as its square root is a symmetric distance measure which is exactly what is required in the MMR measure.

5 DIVERSITY-AWARE EVALUATION METRIC

To be able to evaluate the effectiveness of our result diversity approach, an evaluation metric that takes into consideration both relevance of results as well as their diversity must be used. There is a wealth of work on diversity-aware evaluation metrics for IR systems such as (Allan et al., 2003; Clarke et al., 2008; Zhai et al., 2003). We adopt a similar strategy and propose a novel evaluation metric that takes into consideration both aspects we are concerned with here, namely relevance and diversity, to evaluate a result set for a given query.

We introduce an adjustment to the Discounted Cumulative Gain (DCG) (Jrvelin and Keklinen, 2002) metric by adding a component that takes into consideration the novelty of a certain result, which reflects result diversity in a given result set.

More formally, given a particular result set (a result ordering) of p results, the diversity-aware DCG, which we coin *DIV-DCG* is computed as follows:

$$DCG_p = rel_1 + nov_1 + \sum_{i=2}^{p} (\frac{rel_i}{log_2(i)} + nov_i) \tag{7}$$

where rel_i is the relevance score of the result at position i and nov_i is its novelty.

5.1 Resource-based Novelty

Concerning our first two diversity notions: the resource-based and term-based, the novelty of a result at position i can be computed as follows:

$$nov_i = \frac{\#unseen_i}{\#variables} \tag{8}$$

where $\#unseen_i$ is the number of resources that are bound to variables in result at position i that have not yet been seen, and *#variables* is the total number of variables in the query. Our goal is to diversify the results with respect to the variable bindings.

5.2 Text-based Novelty

The computation of the text-based novelty metric is very similar in spirit to the resource-based one. The only difference is that in the case of text-based diversity, our goal is to diversify the results with respect to their text snippets. To be able to quantify this, we measure for each result, the amount of new keywords that this result contributes to the set of keywords of the previously ranked results. More precisely, the text-based novelty can be computed as follows:

$$nov_i = \frac{|keywords_i \setminus (keywords_i \cap (\cup_{j=1}^{i-1} keywords_j))|}{|keywords_i|} \tag{9}$$

where $keywords_i$ is the set of the keywords associated with the subgraph at position i, $\cup_{j=1}^{i-1} keywords_j$ is the set of all the keywords seen so far (up to subgraph $i-1$), and $|keywords_i|$ is the number of keywords in the set $keywords_i$.

6 EVALUATION

6.1 Setup

In order to evaluate our greedy diversity-aware reranking algorithm, we constructed a benchmark of

Table 3: Average $DIV-NDCG_{10}$ of the training queries for different values of λ.

Notion	$\lambda=0.1$	$\lambda=0.2$	$\lambda=0.3$	$\lambda=0.4$	$\lambda=0.5$	$\lambda=0.6$	$\lambda=0.7$	$\lambda=0.8$	$\lambda=0.9$
Resouce-based	**0.798**	0.789	0.780	0.769	0.762	0.758	0.751	0.740	0.727
Term-based	**0.798**	0.785	0.770	0.759	0.752	0.747	0.741	0.731	0.724
Text-based	**0.932**	0.931	0.928	0.923	0.918	0.915	0.906	0.892	0.875

100 queries over DBPedia, which can be broadly divided into 4 categories: structured, keyword-augmented, requiring relaxation, and keyword-augmented and requiring relaxation.

Our query benchmark was used in order to tune the weighting parameter of MMR (λ in Equation 5), that is used to trade-off relevance and diversity.

We needed to gather a relevance assessment for each result to be able to compute rel_i. To gather these relevance assessments, we relied on the crowdsourcing platform CrowdFlower as follows. For each one of our three diversity notions, each query was run 10 times with λ ranging from 0.1 to 1 (i.e., no diversity) and the top-10 results were retrieved.

The inter-rater agreement reported by CrowdFlower was *67%*. In addition, we computed the Fleiss Kappa Coefficient as another measure of agreement which is more reliable measure than that of CrowdFlower as it takes into consideration agreement by chance. We obtained a Kappa Coefficient of *40%* which can interpreted as "Fair Agreement" (Fleiss, 1971).

Table 4: Average $DIV-NDCG_{10}$ of the test queries.

Notion	Diversified Result Set	Non-diversified Result Set
Resource-based	0.79	0.74
Term-based	0.81	0.74
Text-based	0.91	0.68

To measure the efficiency of our algorithm, we calculate the average execution time for each notion. For the Resource-based and Term-based notions, the averages are 5.7 s and 5.9 s respectively. For the Text-based, it is 26 s.

6.2 Experiments

6.2.1 Parameter Tuning

The main parameter in our diversity-aware re-ranking algorithm is the weighting parameter λ which trades-off relevance and diversity. To be able to set this parameter, we divided our query benchmark into a training set consisting of 80 queries and computed the *normalized* $DIV-DCG_{10}$ for each query in the training set varying the value of λ from 0.1 to 0.9. The normalized $DIV-DCG_{10}$ or $DIV-NDCG_{10}$ is computed by dividing the $DIV-DCG_{10}$ by the *ideal* $DIV-DCG_{10}$. To be able to compute the ideal $DIV-DCG_{10}$, we re-ranked the results using a greedy approach. The new ordering pushes the results with the best combination of diversity and relevance gain to the top. Table 3 shows the average $DIV-NDCG_{10}$ for our three notions of diversity for different values of λ. For our three notions, the best value for λ is 0.1, which means we should give 90% importance to diversity over relevance in our re-ranking algorithm in order to get the best $DIV-NDCG_{10}$ possible.

6.2.2 Comparison of Various Notions of Diversity

Given the optimal values of the parameter λ that were set based on the 80 training queries in our benchmark, we computed the average $DIV-NDCG_{10}$ for the 20 test queries for each notion of diversity, as well as for the cases when no diversity was employed. The summary of our findings over the 20 queries is shown in Table 4. As can be seen from the table, the average $DIV-NDCG_{10}$ for all notions is significantly larger than those where no diversification of results took place.

Next we show some qualitative results that highlight the importance of result diversity for RDF search and the difference between the various notions of diversity we discussed here.

Resource-based Diversity. Consider the query *"?film1 director ?x; ?film2 starring ?y; ?x spouse ?y"* whose corresponding information need is "Give me the name of a director whose partner is an actor (or actress) and the name of a movie for each". Table 5 shows the top-5 results with and without diversification, with λ set to 0.1. The table shows that without diversification, the results are dominated by two resources, namely *Madonna* and *Sean_Penn* with different combinations of movies they acted in or directed, but when diversification is involved, we get a set of different director-actor pairs as shown in the second column of Table 5.

Term-based Diversity. Consider one of the test queries: *"?m director ?x [Disney]; ?m starring ?y"*. The information need is: "Give me the name of a a movie, its director, and its star, preferably a Disney movie". The top-5 results with the diversity parameter λ set to 0.1 are shown in Table 6. In the non-

Table 5: Top-5 results for the query: *"?film1 director ?x; ?film2 starring ?y; ?x spouse ?y"* with and without diversity.

No Diversity	Resource-based Diversity
W.E. director Madonna Mystic_River starring Sean_Penn Madonna spouse Sean_Penn	W.E. director Madonna Mystic_River starring Sean_Penn Madonna spouse Sean_Penn
Secretprojectrevolution director Madonna Mystic_River starring Sean_Penn Madonna spouse Sean_Penn	Citizen_Kane director Orson_Welles Cover_Girl starring Rita_Hayworth Orson_Welles spouse Rita_Hayworth
W.E. director Madonna The_Tree_of_Life starring Sean_Penn Madonna spouse Sean_Penn	Henry_V director Laurence_Olivier Ship_of_Fools starring Vivien_Leigh Laurence_Olivier spouse Vivien_Leigh
W.E. director Madonna Dead_Man_Walking starring Sean_Penn Madonna spouse Sean_Penn	That_Thing_You_Do! director Tom_Hanks Jingle_All_the_Way starring Rita_Wilson Tom_Hanks spouse Rita_Wilson
W.E. director Madonna This_Must_Be_the_Place starring Sean_Penn Madonna spouse Sean_Penn	Then_She_Found_Me director Helen_Hunt The_Simpsons_Movie starring Hank_Azaria Helen_Hunt spouse Hank_Azaria

Table 6: Top-5 results for the query *"?m director ?x [Disney]; ?m starring ?y"* with and without diversity.

No Diversity	Term-based Diversity
Aladdin_(1992_Disney_film) director John_Musker Aladdin_(1992_Disney_film) starring Robin_Williams	Aladdin_(1992_Disney_film) director John_Musker Aladdin_(1992_Disney_film) starring Robin_Williams
Aladdin_(1992_Disney_film) director Ron_Clements Aladdin_(1992_Disney_film) starring Robin_Williams	102_Dalmatians director Kevin_Lima 102_Dalmatians starring Glenn_Close
Aladdin_(1992_Disney_film) director John_Musker Aladdin_(1992_Disney_film) starring Frank_Welker	Cars_2 director John_Lasseter Cars_2 starring Michael_Caine
Aladdin_(1992_Disney_film) director Ron_Clements Aladdin_(1992_Disney_film) starring Frank_Welker	Leroy_&_Stitch director Tony_Craig Leroy_&_Stitch starring Tara_Strong
Aladdin_(1992_Disney_film) director John_Musker Aladdin_(1992_Disney_film) starring Gilbert_Gottfried	Snow_White_and_the_Seven_Dwarfs director Wilfred_Jackson Snow_White_and_the_Seven_Dwarfs starring Pinto_Colvig

Table 7: Top-5 results for the query *"?m distributor Columbia_Pictures"* with and without diversity.

No Diversity	Text-based Diversity
Spider-Man_2 distributor Columbia_Pictures	Spider-Man_2 distributor Columbia_Pictures
Close_Encounters distributor Columbia_Pictures	Sharkboy_&_Lavagirl distributor Columbia_Pictures
A_Clockwork_Orange distributor Columbia_Pictures	Jungle_Menace distributor Columbia_Pictures
Spider-Man_3 distributor Columbia_Pictures	Quantum_of_Solace distributor Columbia_Pictures
Lawrence_of_Arabia distributor Columbia_Pictures	The_Da_Vinci_Code distributor Columbia_Pictures

diversified set of results, we have one popular Disney movie, *Aladdin*, that is repeated five times. On the other hand, the diversified set of results contains five different *Disney* movies. Note that, if we had used the resource-based diversity notion to diversify the results of this query, we would have ended up with 5 different movies which would not have been necessarily Disney movies.

Text-based Diversity. Consider one of the test queries: *"?m distributor Columbia_Pictures"*. The corresponding information need is: "Give me the name of a film distributed by the Columbia Pictures company". The top-5 results for both no diversity, and diversity notion 3 with the diversity parameter λ set to 0.1 are shown in Table 7. While the non-diversified set contains different movies, these movies in fact are not very diverse. Unlike the non-diversified set, the diversified set of results contains movies that have totally different genres, locations, casts, and plots. This indirect level of diversity cannot be captured using our first two diversity notions.

7 CONCLUSION

In this paper, we proposed a framework to diversify the results of triple-pattern queries over RDF datasets. We provided the first formal definition of result diversity in the context of RDF search based on three different notions. We also developed the first diversity-aware ranking model for RDF search, which is based on the Maximal Marginal Relevance. Finally, we designed a new diversity-aware evaluation metric for RDF search based on the Discounted Cummulative Gain and used it on a benchmark of 100 queries over

DBPedia. Our experimental evaluation highlights the importance of result diversity in the context of RDF search and the flexibility of our approach and notions of diversity in capturing different aspects of user queries.

In future work, we plan to carry out more experiments on other RDF datasets to further validate our results. We also plan to explore other notions of diversity such as an ontology-based diversity notion where results can be diversified based on resource types for instance. Finally, we plan to study other relevance measures and to investigate other metrics for computing result diversity instead of the language modeling approach we adopted in this paper.

ACKNOWLEDGEMENT

We would like to thank the American University of Beirut's research board (URB) for funding our research.

REFERENCES

Agrawal, R., Gollapudi, S., Halverson, A., and Ieong, S. (2009). Diversifying search results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 5–14, New York, NY, USA. ACM.

Allan, J., Wade, C., and Bolivar, A. (2003). Retrieval and novelty detection at the sentence level. In *SIGIR*, pages 314–321.

Auer, S., Bizer, C., Cyganiak, R., Kobilarov, G., Lehmann, J., and Ives, Z. (2007). Dbpedia: A nucleus for a web of open data. In *ISWC/ASWC*.

Carbonell, J. and Goldstein, J. (1998). The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *SIGIR*.

Chaudhuri, S., Das, G., Hristidis, V., and Weikum, G. (2006). Probabilistic information retrieval approach for ranking of database query results. *SIGMOD Record*, 35(4).

Chen, H. and Karger, D. R. (2006). Less is more: probabilistic models for retrieving fewer relevant documents. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, pages 429–436, New York, NY, USA. ACM.

Chen, Z. and Li, T. (2007). Addressing diverse user preferences in SQL-query-result navigation. In *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, SIGMOD '07, pages 641–652, New York, NY, USA. ACM.

Clarke, C. L., Kolla, M., Cormack, G. V., Vechtomova, O., Ashkan, A., Büttcher, S., and MacKinnon, I. (2008). Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 659–666, New York, NY, USA. ACM.

Dali, L., Fortuna, B., Tran Duc, T., and Mladenic, D. (2012). Query-independent learning to rank for rdf entity search. In *ESWC*, pages 484–498.

Elbassuoni, S., Ramanath, M., Schenkel, R., Sydow, M., and Weikum, G. (2009). Language-model-based ranking for queries on RDF-graphs. In *CIKM*.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378 – 382.

Gollapudi, S. and Sharma, A. (2009). An axiomatic approach for result diversification. In *Proceedings of the 18th international conference on World wide web*, WWW '09, pages 381–390, New York, NY, USA. ACM.

Jrvelin, K. and Keklinen, J. (2002). Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, pages 422–446.

Kasneci, G., Suchanek, F. M., Ifrim, G., Ramanath, M., and Weikum, G. (2008). Naga: Searching and ranking knowledge. In *ICDE*.

Lin., J. (1991). Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, pages 145–151.

RDF (2004). W3c: Resource description framework (rdf). www.w3.org/RDF/.

SPARQL (2008). W3c: Sparql query language for rdf. www.w3.org/TR/rdf-sparql-query/.

Suchanek, F. M., Kasneci, G., and Weikum, G. (2008). Yago: A large ontology from wikipedia and wordnet. *J. Web Sem.*, 6(3).

Vee, E., Srivastava, U., Shanmugasundaram, J., Bhat, P., and Yahia, S. A. (2008). Efficient Computation of Diverse Query Results. In *Proceedings of the 2008 IEEE 24th International Conference on Data Engineering*, pages 228–236, Washington, DC, USA. IEEE Computer Society.

Zhai, C. X., Cohen, W. W., and Lafferty, J. (2003). Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, SIGIR '03, pages 10–17, New York, NY, USA. ACM.