max planck institut informatik

1866 AUB American University of Beirut الجامعة الأميركية في بيروت

Effective Searching of RDF Knowledge Graphs

Hiba Arnaout[1] & Shady Elbassuoni[2]

1 Max Planck Institute for Informatics
2 American University of Beirut

RDF Knowledge Graphs(KG)

Formal representation

RDF KG

DBpedia yago select knowledge WIKIDATA Freebase

```
<Subject; Predicate; Object>
<The Godfather; director; Francis Ford Coppola>
<The Godfather; producer; Albert S. Ruddy>
<The Godfather; starring; Al Pacino>
<The Godfather; distributor; Paramount Pictures>
```

RDF triples

Resource Description Framework

Querying..
SPARQL
Triple-pattern queries

```
<?m; director; Francis Ford Coppola; ?m; starring; Al Pacino>
```

Challenges

Input:
Triple-pattern Query

?

Output:
List of results

Challenges (continued)

1 **Triplifying everything**

Text will always have more information

Input:
Triple-pattern Query

<?m; director; ?x; ?m; starring; ?y>

Output:
List of results

Query need: "List of movies about *weddings* that took place in *New York*, their *directors* and their *stars*"

?m	?x	?y
Suspect	Peter Yates	Cher
Book Club	Bill Holderman	Jane Fonda
Suspect	Peter Yates	John Mahoney
Prom	Joe Nussbaum	Yin Chang

Random movies with their directors and stars

Challenges (continued)

1 Triplifying everything

Text will always have more information

Input:
Triple-pattern Query

2 Flexible querying

Result either matches the query or not

Output:
List of results

Query need: "Movies *directed* by *Spielberg*"

```
<Stephen Spielberg; director; ?m>
```

?m
Empty list of results

Challenges (continued)

1 Triplifying everything

Text will always have more information

Input:
Triple-pattern Query

2 **Flexible querying**

Result either matches the query or not

Output:
List of results

Query need: "Movies *directed* by *Spielberg*"

<Stephen Spielberg; director; ?m>

<?m; director; Stephen Spielberg>

?m
Empty list of results

Challenges (continued)

1 Triplifying everything

Text will always have more information

2 Flexible querying

Result either matches the query or not

3 **Result ranking**

Too many results for a given query

Input:

Triple-pattern Query

Output:

List of results

?m	?x
Bordellet	Olge Edge
Marx Reloaded	Jason Barker
The Kid	Charlie Chaplin
Annie Hall	Woody Allen
Kranti	Manoj Kumar
Harlem Nights	Eddie Murphy
2000 results	

Query need: "Movies *directed* by *one of their stars*"

```
<?m; director; ?x; ?m; starring; ?x>
```

Challenges (continued)

1 Triplifying everything

Text will always have more information

2 Flexible querying

Result either matches the query or not

3 Result ranking

Too many results for a given query

4 Result diversity

Homogeneous sets of results

Input:

Triple-pattern Query

Query need: "Publishers of *Stephen King* books"

```
<?b; author; Stephen King; ?b; publisher; ?p>
```

Output:

List of results

?b	?p
Cycle of the Werewolf	New Mexico
Black House	Random House
The Shining	Doubleday
Pet Sematary	Doubleday
Gray Matter	Doubleday
The Ledge	Doubleday

Contributions

- Semantic Web problems using Information Retrieval techniques:
  1. Keyword-extended queries over keyword-extended triples.
  2. Automatic relaxation of queries with zero results.
  3. Ranking model based on statistical machine translation.
  4. Notions for result diversity and a re-ranking algorithm based on Maximal Marginal Relevance (Carbonell J., Goldstein J.,1998).
  5. Diversity-aware evaluation metric: relevance and novelty gain.
- Previous Work: a summarized survey can be found in the paper.

Ranking

Given a query(*possibly extended with keywords*)…..

Goal: produce a list of *top-k results* ranked by relevance.

- Statistical language-modeling-based approach:
  - Statistical machine translation
  - Probability of result generating every triple pattern of the query:
    1. If *NO keyword-extended* query: weights of the triples in the result.
    2. If *keyword-extended* query: weights of the triples AND weights of the triples with respect to the keywords associated.

Ranking (continued)

<?m; director; ?x; ?m; starring; ?x>

NON-RANKED	?m	?x
1	Bordellet	Olge Ege
2	Marx Reloaded	Jason Barker
3	The Almost Guys	Eric Fleming
4	Kranti	Manoj Kumar
5	Purab Aur Paschim	Manoj Kumar

RANKED	?m	?x
1	None But the Brave	Frank Sinatra
2	Cruel Summer	Kanye West
3	The Bond	Charlie Chaplin
4	A Dog's Life	Charlie Chaplin
5	Citizen Kane	Orson Welles

Ranking (continued)

<?m; director; ?x; ?m; starring; ?x>

+ []

RANKED	?m	?x
1	None But the Brave	Frank Sinatra
2	Cruel Summer	Kanye West
3	The Bond	Charlie Chaplin
4	A Dog's Life	Charlie Chaplin
5	Citizen Kane	Orson Welles

+ [novel]

RANKED	?m	?x
1	Lady from Shanghai	Orson Welles
2	The Trial	Orson Welles
3	Journey into Fear	Orson Welles
4	The Prince of Tides	Barbra Streisand
5	Firefox	Clint Eastwood

Relaxation

1. Reasons of query failure
   - three possible cases
2. Move constants causing the problem to the keywords space.
3. Replace them with variables.
4. Produce and execute the set of relaxed queries.
5. Rank each result using our ranking model:
   - while giving higher scores for results produced by more than one of the relaxed queries.

Relaxation (continued)

<?x; starring; Frank Senatra> <?x; producer; ?p>

becomes…

<?x; starring; ?y [frank senatra]; ?x; producer; ?p>

RELAXED THEN RANKED	?x	?y	?p
1	None But the Brave	Frank Sinatra	Frank Sinatra
2	On the Town	Frank Sinatra	Arthur Freed
3	The Amazing Mr. Bickford	Frank Zappa	Frank Zappa
4	From Here to Eternity	Frank Sinatra	Buddy Adler
5	Take Me Out to the Ball Game	Frank Sinatra	Arthur Freed

Diversity

Goal: Re-score all the results by BOTH relevancy and novelty.

- MMR(Maximal Marginal Relevance)*
- Measure how much every result in A will contribute:
  - To the relevance of B: using our ranking model.
  - To the diversity of B:
    - ***What is the notion of diversity in the setting of RDF?***
    - Distance between a candidate result, and all the other results that made it to B :
      1. **Knowledge-graph-based diversity**
      2. **Query-based diversity**
      3. **Text-based diversity**

*Carbonell J., Goldstein J.(1998) The Use of MMR, diversity-based reranking for reordering documents and producing summaries

Diversity: knowledge-graph-based

<?b; author; Stephen King; ?b; publisher; ?p>

NON-DIVERSED (RELEVANCE)	?b	?p
1	Cycle… Werewolf	New Mexico
2	Black House	Random House
3	The Shining	Doubleday
4	Pet Sematary	Doubleday
5	Gray Matter	Doubleday

DIVERSIFIED	?b	?p
1	Cycle… Werewolf	New Mexico
2	Black House	Random House
3	The Shining	Doubleday
4	My Pretty Pony	Alfred_A._Knopf
5	The Dead Zone	Viking Press

Diversity: query-based

<?x; starring; ?y[chaplin]; ?x; director; ?y>

DIVERSIFIED (KG-based)	?x	?y
1	City Lights	Charlie Chaplin
2	Braveheart	Mel Gipson
3	Pinched	Harold Lloyd
4	Annie Hall	Woody Allen
5	Boxes	Jane Birkin

DIVERSIFIED (Q-based)	?x	?y
1	City Lights	Charlie Chaplin
2	Limelight	Charlie Chaplin
3	Braveheart	Mel Gipson
4	Pay Day	Charlie Chaplin
5	The Tramp	Charlie Chaplin

Diversity: text-based

<?x; distributor; Warner Bros>

NON-DIVERSIFIED (RELEVANCE)	?x
1	The Dark Night
2	Tom and Jerry
3	Blade Runner
4	A Space Odyssey
5	Dragnet
6	Slumdog Millionaire
7	Citizen Kane
8	Batman Begins
9	Inception
10	Batman

DIVERSIFIED (T-based)	?x
1	The Dark Night
2	Sweeney Todd
3	The Coo-Coo Nut Grove
4	Dragnet
5	A Space Odyssey
6	Dive Bomber
7	Blade Runner
8	Battlefield Earth
9	Mindscape
10	Adventures of Road Runner

Experiments

1. Construct a Benchmark of 139 queries (available online):
   https://github.com/HibaArnaout/sup_material
2. CrowdFlower, a Crowdsourcing platform.
3. Evaluate Ranking model(including relaxation)
   a) *Non-ranked* sets, only matching results.
   b) Same ranking model, but *different triple-weight* computation.
   c) *Language-modeling ranking approach*(Elbassuoni S. et al, 2009).

Experiments (continued)

4. Every worker is asked to choose *between* set of results X and Y, or *same*:

   Won over a) **72%**, over b) **63%**, and **Tied** with c) **94%**.

5. Evaluate Diversity model against our non-diversified model:
   - Knowledge-graph-based: **Won 50%** and **Tied 44%.**
   - Query-based: **Won 54%** and **Tied 41%.**
   - Text-based: **Won 38%** and **Tied 56%.**

Conclusion

1. Extend the knowledge graph with *keywords* and allow *keyword-extended queries*.
2. *Ranking model* based on statistical machine translation.
3. Automatic *query relaxation* that preserve the original query intention.
4. Define three notions for *diversity*, and adapt a re-ranking algorithm based on MMR.
5. Create a *benchmark* of different query categories to evaluate different problems.
6. Propose a *diversity-aware evaluation metric*, an updated version of the popular normalized discounted cumulative gain.

Thank you!

(Harnaout@mpi-inf.mpg.de)

References at: https://bit.ly/2ysrYf9