Ranking Entities for Web Queries Through Text and Knowledge

Speaker: Shih-Han Lo
Advisor: Professor Jia-Ling Koh
Author: Michael Schuhmacher, Laura Dietz, Simone Paolo Ponzetto
Date: 2016/12/06
Source: CIKM '15

Outline

- **Introduction**
- Method
- Experiment
- Conclusion

Introduction

Motivation

- Web search engine research shows an increasing interest in going beyond words.
- Entity ranking has recently attracted a lot of attention from researches.

Introduction

- The contributions:
  - New task: A new variant of ad-hoc entity search.
  - New evaluation dataset: We create an gold standard from queries.
  - Experimental study: We use ground truth judgements from our dataset.

Introduction

Entity ranking task

- Given: A user-provided text query.
- Provided: A background corpus with query-relevant documents that includes entity link annotations.
- Goal: Rank the entities by relevance.

Introduction

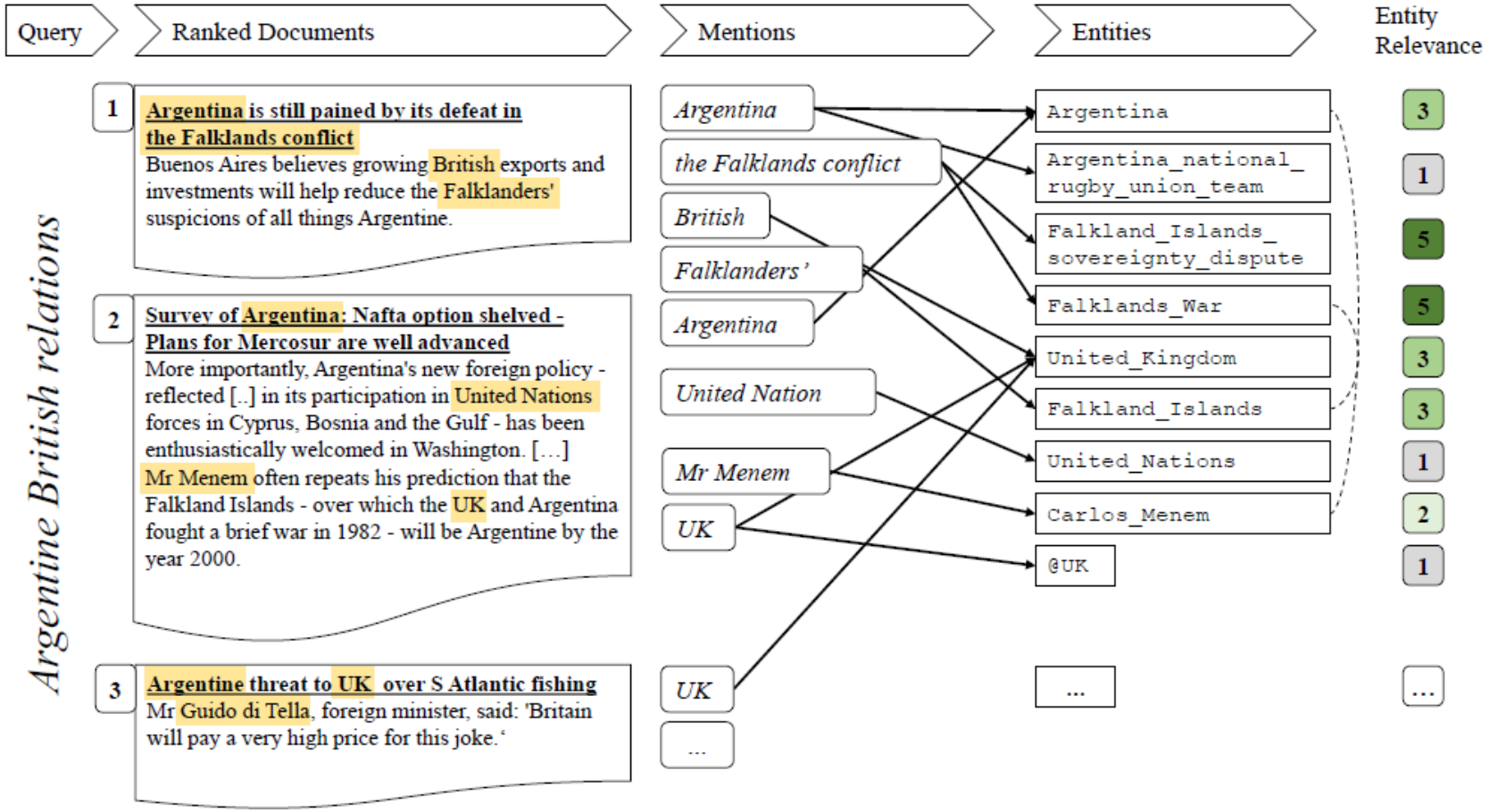


Figure 1: An example of ranking entities for a Web query. Solid arrows represent entity links, dotted lines knowledge base relations.

Outline

- Introduction
- **Method**
- Experiment
- Conclusion

Method

Learning to Rank Entities for Web Queries

- Train different supervised models on labeled data.
- Employ different learning-to-rank algorithms:
  - Ranking Support Vector Machine (SVM)
  - An algorithm to optimize an underlying retrieval metric

Method

Document Retrieval and Candidate Pool

- We study two retrievals for two datasets:
  - Robust04: EQFE (entity-aware)
  - ClueWeb12: Sequential Dependency Model (SDM) and FACC1 dataset
- Entity links in high ranked documents are used to build a pool of candidate entities.

Method

Mention Features

- Mention Frequency (MenFrqIdf)
  - We use TF-IDF to weight mention counts (compare with MenFrq)

$$\text{MenFrqIdf}(e) = \text{tf}_q(e) \log \frac{N}{\text{df}(e)}$$

Method

Query-Mention Features

- String Edit Distance (SED)
  - To cover basic morpho-syntactic similarity.
- GloVe (Glo) / JoBimText (Jo)
  - To model word meanings based on their global frequency of co-occurrences in large text corpora.

Method

Query-Entity Features

- Entity-set Comparison Using Dbpedia
  - First run the queries through TagMe.
  - We incorporate 2 methods between both sets:
    - Direct Match (QEnt)
    - Path (QEntEntSim)

Method

Query-Entity Features

- Knowledge Base Retrieval Using Wikipedia
  - WikiBoolean
    - Test if at least one query keyword is found within the Wikipedia article of the candidate entity.
  - WikiSDM
    - Use the query to retrieve 1,000 Wikipedia articles.
    - Re-rank entities that have a high mention frequency.

Method

Entity-Entity Features

- We use QEntEntSim to compute a similarity matrix.
- Extend the ranking SVM with a custom kernel (consists of the following):
  - Linear kernel
  - Semantic smoothing kernel

Outline

- Introduction
- Method
- **Experiment**
- Conclusion

Experiment

REWQ datasets

- REWQ-Robust
  - Build on the TREC Robust 2004 data set.
  - Relevance score: 5-level Likert scale (1 to 5)
- REWQ-ClueWeb
  - Build upon TREC 2013/2014 with the ClueWeb12 corpus.
  - Binary relevance (0 or 1)

Experiment

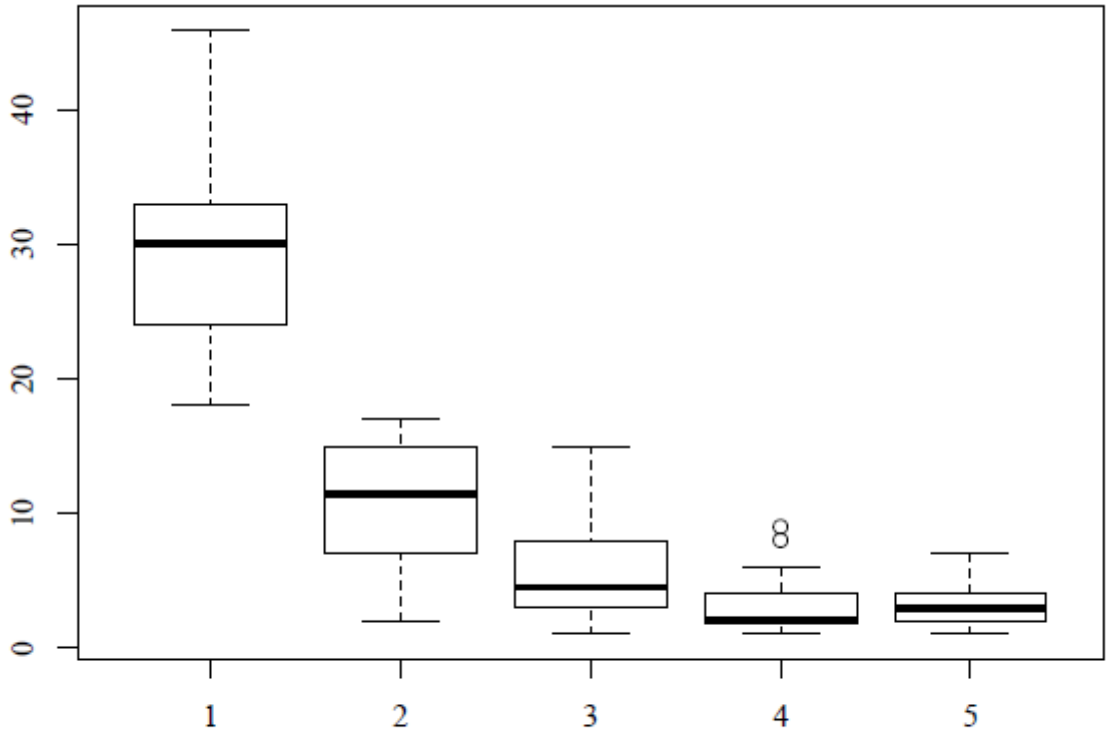


Figure 2: Boxplot for ground truth labels of REWQ-Robust (rounded to a 5-level scale) for 25 query with 50 entities each.

Experiment

- Evaluation metrics:
  - Normalized Discounted Cumulative Gain (NDCG)
  - Mean Average Precision (MAP, for REWQ-ClueWeb)

- Experimental settings:
  - Mention Frequency (MenFreqIdf)

$$\mathrm{tf}_q(e) = \mathrm{TRR}_q(e) = \sum_{d \in D(q)} \sum_{m} \frac{1}{rank_{e,m,d}}$$

  - Wikipedia Fulltext Index (WikiSDM)
  - Wikipedia PageRank (WikiPR)

Experiment

Table 1: Evaluation results for REWQ Robust dataset. We report differences w.r.t. the best performing reference method (here WikiSDM), statistically significant improvements are denoted with † (paired t-test p-value ≤ 0.05).

Method	ndcg	Δ%	ndcg10	Δ%
RankLib	0.936	†3.7	0.817	†11.6
SVM (w/ SK)	0.926	†2.6	0.804	†9.7
SVM (w/o SK)	0.923	2.2	0.796	†8.7
WikiSDM	0.903	0.0	0.733	0.0
MenFrqIdf	0.885	-2.0	0.694	-5.3
WikiPR	0.778	-13.8	0.440	-40.0

Table 2: Evaluation results for REWQ ClueWeb12 dataset. We report differences w.r.t. the best performing reference method (here MenFrqIdf), statistically significant improvements are denoted with † (paired t-test p-value ≤ 0.05).

	map	Δ%	ndcg	Δ%	ndcg10	Δ%
RankLib	0.328	†9.0	0.572	†3.4	0.710	†10.0
SVM (w/ SK)	0.278	-7.8	0.545	-1.6	0.646	0.1
SVM (w/o SK)	0.308	2.2	0.563	1.6	0.675	4.4
MenFrqIdf	0.301	0.0	0.554	0.0	0.646	0.0
WikiSDM	0.234	-22.3	0.515	-7.0	0.613	-5.1
WikiPR	0.075	-75.1	0.328	-40.8	0.126	-80.5

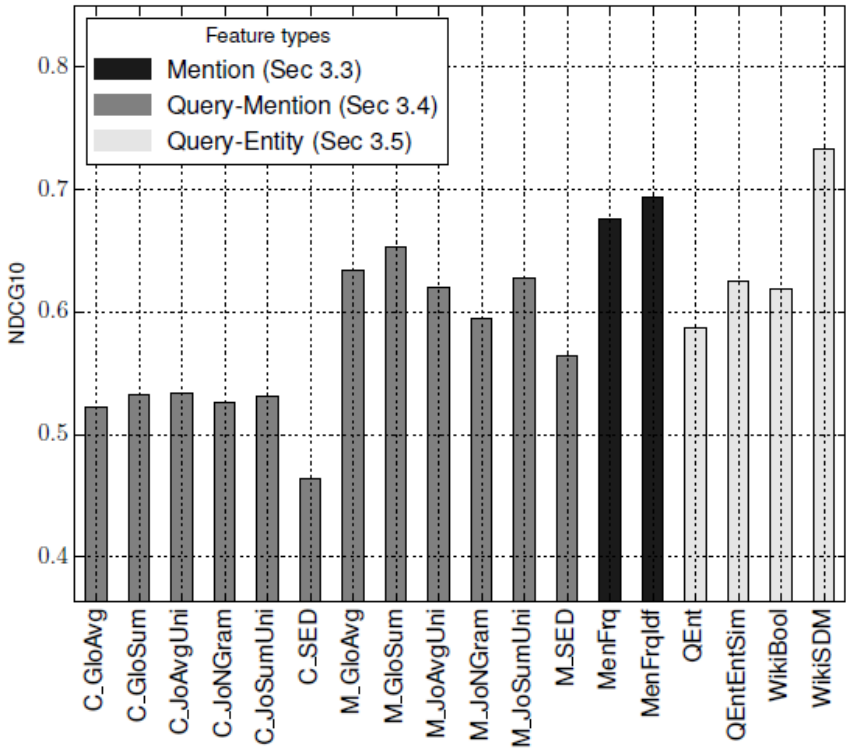
Experiment

Table 3: List of all REWQ Robust04 queries sorted by average ground truth scores for top 3 entities (*gt*). Also showing the NDCG@10 score (*ndcg*) and the top-3 entities as retrieved by the RankLib coordinate ascent method (using the complete set of features). Ground truth values in brackets; abbreviated entities are denoted with *.

gt	ndcg	query	top-1 entity	top-2 entity	top-3 entity
5.0	.895	schengen agreement	Schengen_Agreement (5)	Schengen_Area (5)	Schengen_Inform._Sys. (5)
5.0	.894	magnetic levitation maglev	Maglev (5)	Shanghai_Maglev_Train (5)	Transrapid (5)
4.8	.931	territorial waters dispute	UN_Law_of_Sea* (4.3)	Aegean_dispute (5)	Kuril_Islands_dispute (5)
4.7	.865	el nino	El_Nino-Southern_Oscillation (5)	Pacific_Ocean (4)	La_Nina (5)
4.3	.942	ferry sinkings	MS_Estonia (5)	MS_Herald_of_Free_Enterprise (5)	Silja_Line (3)
4.3	.927	in vitro fertilization	In_vitro_fertilisation (5)	Fertility_clinic (5)	Shan_Ratnam (3)
4.3	.879	poliomyelitis and post polio	Poliomyelitis (4.5)	Polio_vaccine (4.5)	Jonas_Salk (4)
4.3	.787	osteoporosis	Hormone_therapy* (5)	Estrogen (5)	Nutrition (3)
4.0	.803	industrial espionage	Volkswagen (4)	General_Motors (4)	Opel (4)
3.8	.863	polygamy polyandry polygyny	Al-Arqam (4.5)	Code_of_Personal_Status_(Tunisia) (4.5)	Tunisia (2.5)
3.3	.769	amazon rain forest	Amazon_rainforest (5)	Manaus (2)	Amazon_Basin (3)
3.3	.748	argentine british relations	Foreign_relations_of_Argent. (4)	Argentina_national_rugby_team* (1)	Falklands_War (5)
3.2	.751	antarctica exploration	South_Pole (3.3)	Antarctica (4)	Antarctic_ecozone (2.3)
2.8	.860	supercritical fluids	Supercritical_fluid (5)	Euler_equations_(fluid_dyn)* (2.5)	Biodegradation (1)
2.8	.760	computer viruses	Michelangelo_(cmpt_virus)* (5)	Personal_computer (2)	Computer_industry (1.5)
2.7	.836	lyme disease	Lyme_disease (5)	Centers_for_Disease_Control* (2)	Old_Lyme,_Connecticut (1)
2.7	.669	falkland petroleum exploration	Falkland_Islands (4.3)	Falklands_War (2.7)	Stanley,_Falkland_Islands (1)
2.5	.739	hydroponics	NASA (4)	Jordan (1)	Mars (2.5)
2.3	.868	killer bee attacks	Africanized_bee (3)	Ceratitis_capitata (2)	San_Diego (2)
2.3	.831	implant dentistry	Dentistry (4)	Cochlear_implant (1)	Uni_of_Med_&_Dent_NJ* (2)
2.3	.749	agent orange exposure	Agent_Orange (5)	Agent_Orange_(band) (1)	Agent_Orange_(album) (1)
2.3	.718	king hussein peace	Hussein_of_Jordan (4.5)	Abdullah_II_of_Jordan (1)	Black_Septmbr_Jordan* (1.5)
2.0	.672	counterfeiting money	Counterfeit (3)	Los_Angeles (1)	Novosibirsk (2)
2.0	.582	unsolicited faxes	Fax (4)	Personal_computer (1)	ISDN* (1)
1.9	.966	agoraphobia	Charles_M._Schulz (3.3)	Snoopy (1.3)	UGM-27_Polaris (1)

Experiment

(a) Robust04 Dataset, 1-5 annotations, KBBridge EL



(b) Clue12 Dataset, binary annotation, FACC1 EL

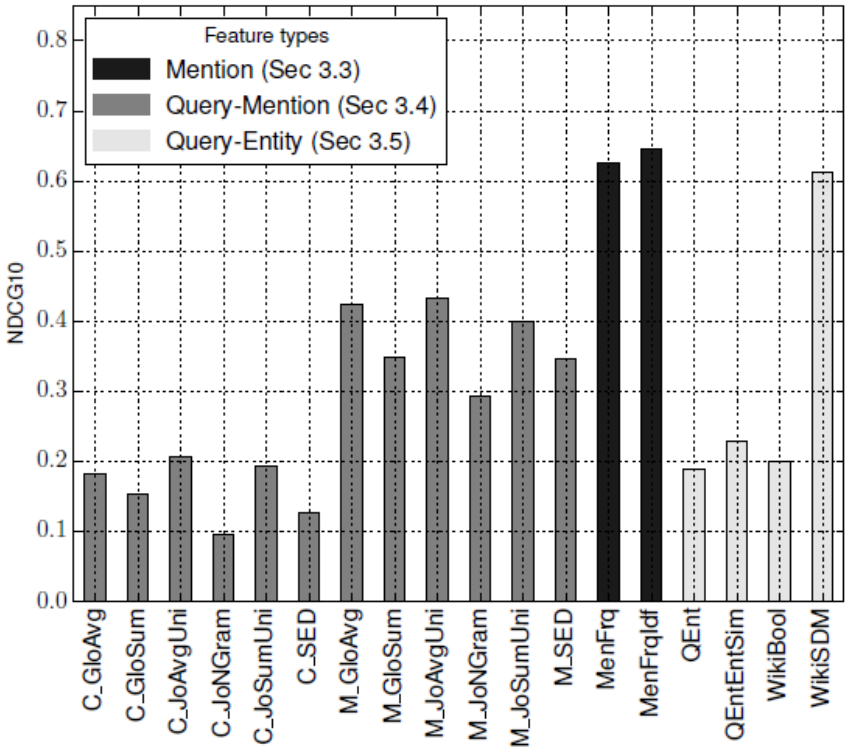


Figure 3: Feature-by-feature analysis for (a) the REWQ-Robust04 and (b) the REWQ ClueWeb12 dataset.

Experiment

Table 4: Feature ablation study REWQ Robust04 dataset. Sorted by difference in NDCG value. P-values (p) from two-sided paired t-test ≤ 0.05 are denoted with †.

w/o	ndcg	Δ%	p	ndcg10	Δ%	p
RankLib All	0.936	-	-	0.817	-	-
MenContext	0.937	0.1	0.68	0.823	0.7	0.56
QEntEnt	0.935	-0.1	0.73	0.824	0.8	0.58
Qent	0.934	-0.2	0.58	0.825	0.9	0.44
Jo	0.934	-0.2	0.53	0.819	0.3	0.85
Context	0.933	-0.3	0.28	0.816	-0.1	0.89
Glo	0.928	-0.8	0.10	0.803	-1.7	0.26
DBpedia	0.927	-1.0	0.06	0.802	-1.9	0.21
WikiBool	0.926	-1.1	0.11	0.809	-1.0	0.56
SED	0.924	†-1.3	0.05	0.794	-2.8	0.09
WikiSdm	0.921	†-1.7	0.03	0.781	†-4.4	0.04
MenFrqIdf	0.917	†-2.1	0.04	0.774	†-5.4	0.05
Wikipedia	0.914	†-2.3	0.01	0.776	†-5.1	0.03

Table 5: Feature ablation study on REWQ ClueWeb12 dataset. Sorted by relative difference (Δ) in MAP value. P-values (p) from two-sided paired t-test ≤ 0.05 are denoted with †.

w/o	map	Δ%	p	ndcg	Δ%	p	ndcg10	Δ%	p
RankLib All	.328	-	-	.572	-	-	.711	-	-
MenContext	.333	1.4	.41	.574	0.3	.55	.714	0.5	.70
Jo	.332	1.0	.55	.573	0.2	.69	.716	0.8	.50
DBpedia	.329	0.1	.92	.572	0.0	.90	.701	-1.4	.26
QEntEnt	.327	-0.4	.48	.572	-0.1	.68	.708	-0.4	.64
Context	.326	-0.6	.49	.570	-0.3	.34	.698	-1.7	.06
Glo	.326	-0.7	.51	.571	-0.3	.46	.698	†-1.7	.05
Qent	.326	-0.8	.63	.571	-0.2	.75	.701	-1.4	.32
SED	.326	-0.8	.35	.571	-0.3	.46	.698	-1.8	.15
WikiSdm	.320	-2.6	.25	.566	-1.1	.26	.693	-2.5	.28
WikiBool	.313	†-4.6	.05	.565	-1.3	.08	.670	†-5.7	.01
Wikipedia	.303	†-7.7	.02	.556	†-2.9	.02	.650	†-8.5	.02
MenFrqIdf	.264	†-19.7	.00	.535	†-6.5	.01	.630	†-11.4	.03

Outline

- Introduction
- Method
- Experiment
- **Conclusion**

Conclusion

- Query-relevant documents with entity links provide a complementary source of information to direct knowledge base retrieval.
- First step that opens up new opportunities for the development of advanced entity-rich search engines.
- Entity ranking for Web queries could play a key role in many high-end applications.

THANK YOU FOR YOUR LISTENING