Multi-step Classification Approaches to Cumulative Citation Recommendation

Krisztian Balog
University of Stavanger

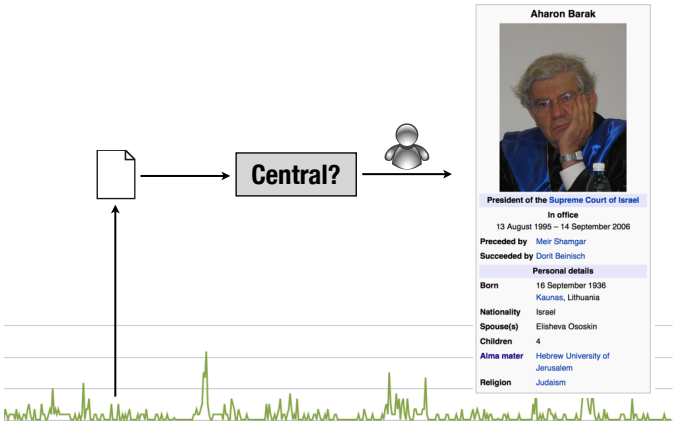
Naimdjon Takhirov, Heri Ramampiaro, Kjetil Nørvåg
Norwegian University of Science and Technology

Open research Areas in Information Retrieval (OAIR) 2013 | Lisbon, Portugal, May 2013

Motivation

- Maintain the accuracy and high quality of knowledge bases
- Develop automated methods to discover (and process) new information as it becomes available

TREC 2012 KBA track



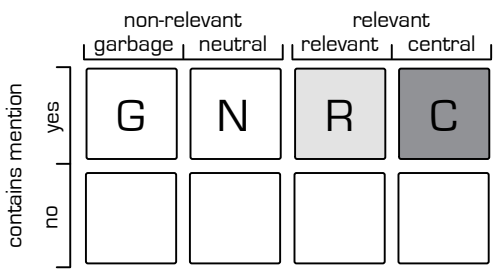
Task
Cumulative Citation Recommendation

- Filter a time-ordered corpus for documents that are highly relevant to a predefined set of entities
  - For each entity, provide a ranked list of documents based on their "citation-worthiness"

Collection and topics

- KBA stream corpus
  - Oct 2011 - Apr 2012
  - Split into training and testing periods
  - Three sources: news, social, linking
  - raw data 8.7TB
  - cleansed version 1.2TB (270G compressed)
  - stream documents uniquely identified by *stream_id*
- Test topics ("target entities")
  - 29 entities from Wikipedia (27 persons, 2 org)
  - uniquely identified by *urlname*

Annotation matrix



Scoring

Target entity: Aharon Barak

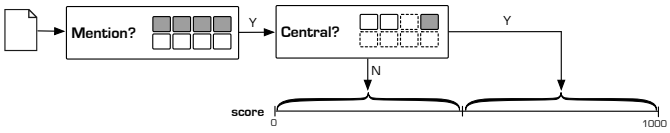
	urlname	stream_id	score	
Positive	Aharon_Barak	1328055120-f6462409e60d2748a0adef82fe68b86d	1000	
	Aharon_Barak	1328057880-79cdee3c9218ec77f6580183cb16e045	500	
	Aharon_Barak	1328057280-80fb850c089caa381a796c34e23d9af8	500	
	Aharon_Barak	1328056560-450983d117c5a7903a3a27c959cc682a	480	
	Aharon_Barak	1328056560-450983d117c5a7903a3a27c959cc682a	450	
	Aharon_Barak	1328056260-684e2f8fc90de6ef949946f5061a91e0	430	
	Aharon_Barak	1328056560-be417475cca57b6557a7d5db0bbc6959	428	
	Aharon_Barak	1328057520-4e92eb721bfbfdfa0b1d9476b1ecb009	428	Cutoff
Negative	Aharon_Barak	1328058660-807e4aaeca58000f6889c31c24712247	380	
	Aharon_Barak	1328060040-7a8c209ad36bbb9c946348996f8c616b	380	
	Aharon_Barak	1328063280-1ac4b6f3a58004d1596d6e42c4746e21	375	
	Aharon_Barak	1328064660-1a0167925256b32d715c1a3a2ee0730c	315	
	Aharon_Barak	1328062980-7324a71469556bcd1f3904ba090ab685	263	

Approach

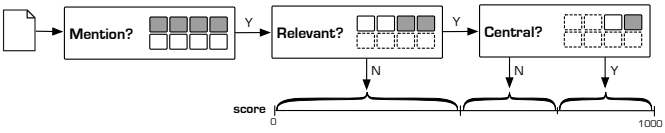
Overview

- "Is this document central for this entity?"
- Binary classification task
- Multi-step approach
  - Classifying every document-entity pair is not feasible
  - First step do decide whether the document contains the entity
  - Subsequent step(s) to decide centrality

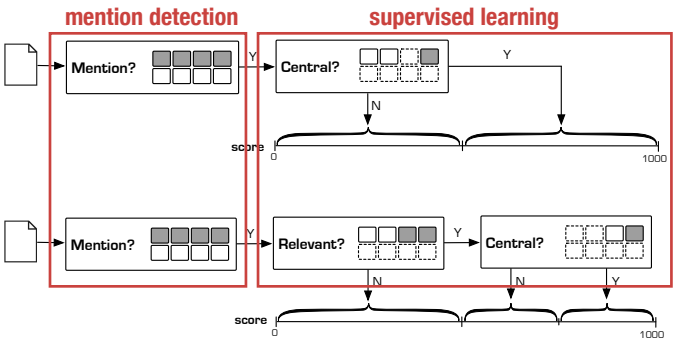
2-step classification



3-step classification



Components



Identifying entity mentions

- Goals
  - High recall
  - Keep false positive rate low
  - Efficiency
- Detection based on known surface forms of the entity
  - urlname (i.e., Wikipedia title)
  - name variants from DBpedia
  - DBpedia-loose: only last names for people
- No disambiguation

Features

1.Document (5)
- Length of document fields (body, title, anchors)
- Type (news/social/linking)

2.Entity (1)
- Number of related entities in KB

3.Document-entity (28)
- Occurrences of entity in document
- Number of related entity mentions
- Similarity between doc and the entity's WP page

Features

4.Temporal (38)
- Wikipedia pageviews
  - Average pageviews
  - Change in pageviews
  - Bursts
- Mentions in document stream
  - Average volume
  - Change in volume
  - Bursts

Results

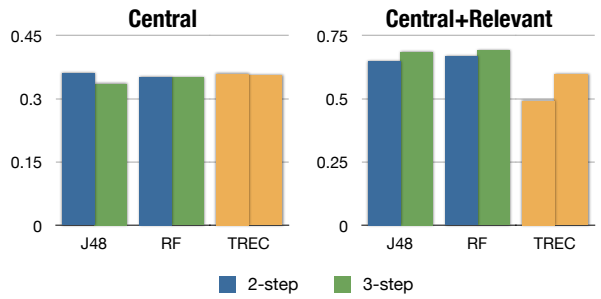
Identifying entity mentions

Results on testing period

Identification	Document-entity pairs	Recall	False positive rate
urlname	41.2K	0.842	0.559
DBpedia	70.4K	0.974	0.701
DBpedia-loose	12.5M	0.994	0.998

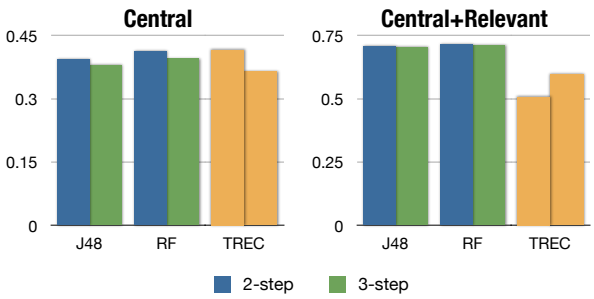
End-to-end task

F1 score, single cutoff



End-to-end task

F1 score, per-entity cutoff



Summary

- Cumulative Citation Recommendation task @TREC 2012 KBA
- Two multi-step classification approaches
- Four groups of features
- Differentiating between relevant and central is difficult

Classification vs. Ranking

[Balog & Ramampiaro, SIGIR'13]

- Approach CCR as a ranking task
- Learning-to-rank
  - Pointwise, pairwise, listwise
- Pointwise LTR outperforms classification approaches using the same set of features

http://krisztianbalog.com/files/sigir2013-kba.pdf

KBAAA
Knowledge Base Acceleration Assessment & Analysis

http://research.idi.ntnu.no/wislab/kbaaa

Questions?

Contact | @krisztianbalog | krisztianbalog.com