A Model for Ranking Entities and Its Application to Wikipedia

Gianluca Demartini,

Claudiu S. Firan, Tereza Iofciu,
Ralf Krestel, Wolfgang Nejdl

Motivation– Entity Search

- Searching the Web you retrieve documents
- "American countries" is an entity search query
- It is easy for a classical search engine to return documents *about* American countries
- The actual countries need to be identified and extracted by the user

- Goal: is to develop a system that can find and return entities
- and not just documents on the Web.

Outline

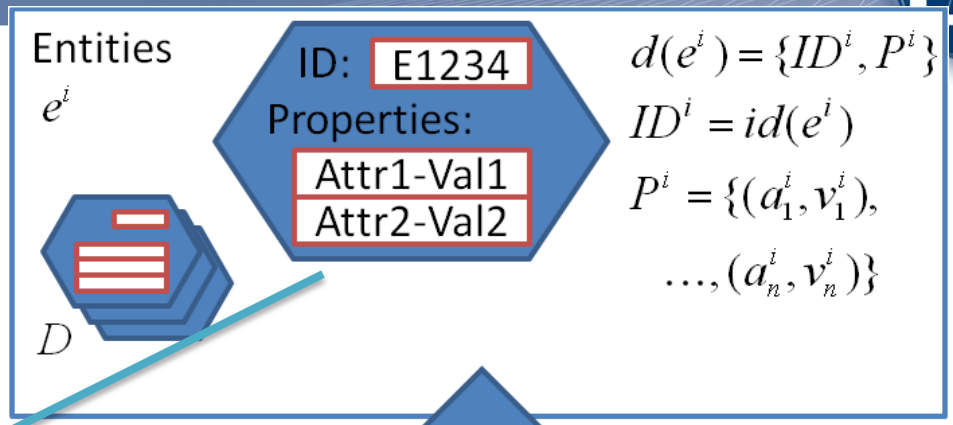
- Formal Model for Entity Ranking
- Application Scenarios
- Wikipedia Setting
- Entity Ranking Algorithms
- Experimental Results
- Demo
- Conclusions

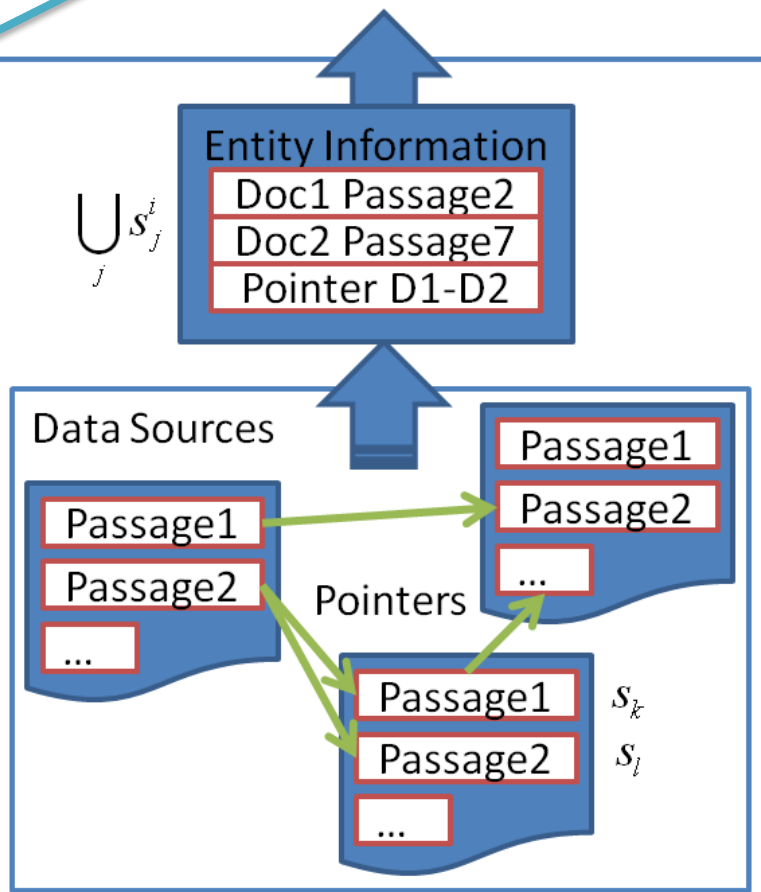
Formal Model for Entity Ranking

- Steps for finding entities:
  - Indexing
    - Entities are identified from data sources
    - Entity descriptions are built
  - Searching
    - User's need is translated into a query
    - IRS extracts from q the *entity need*
    - IRS search the indexed entity description

Formal Model for Entity Ranking

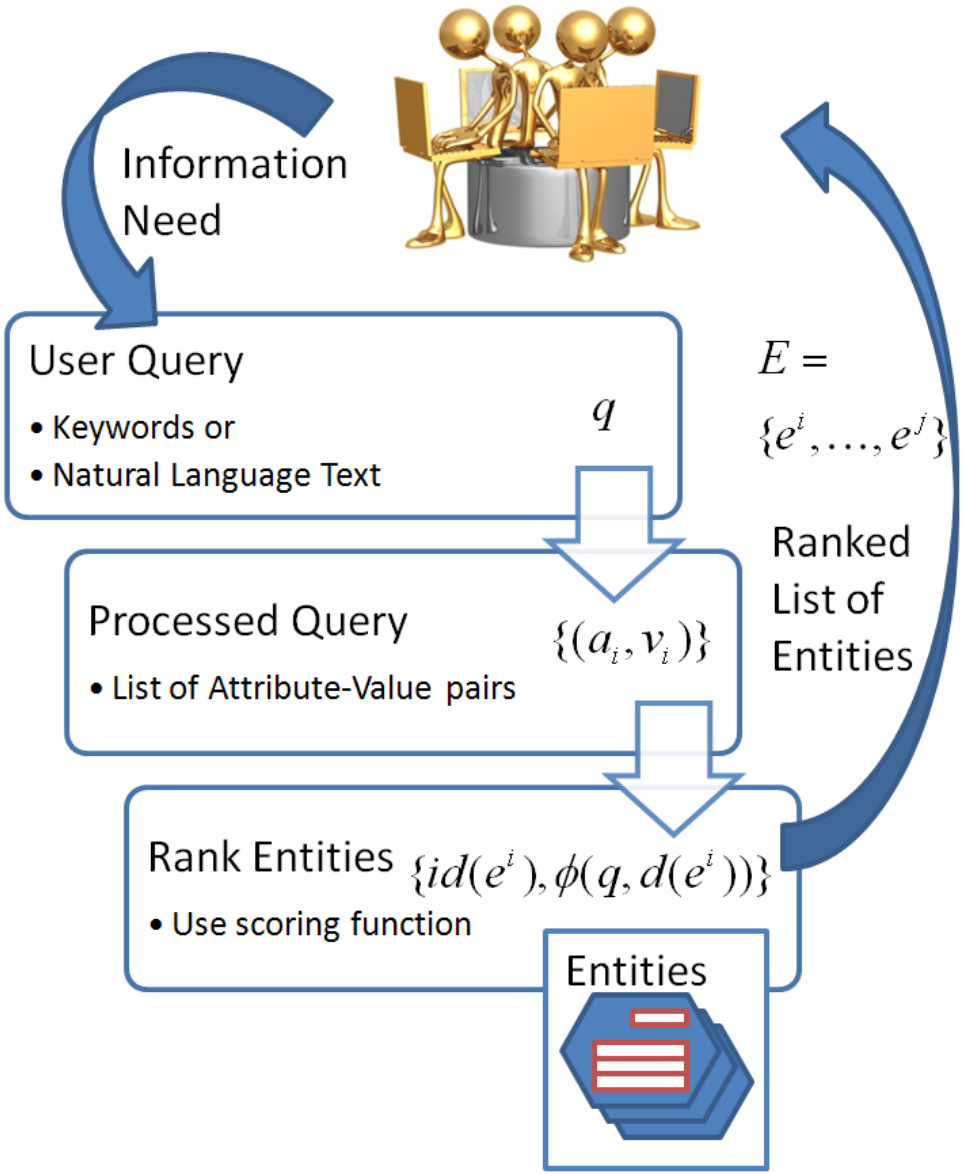
– Indexing
  - Entities
  - Data Sources





"Alexandre Pato"
ID: ap12dH5a
(born in; 1989)
(playing with; acm15hDJ)

Formal Model for Entity Ranking



- Searching
  - Users' Information Need
  - Entity Ranking System

Application Scenarios

- Global Entity Identifiers: OKKAM.org

- Ranking Consumer Products
- Ranking Entities on the Web
- Ranking Entities in Wikipedia

Application to Wikipedia

- INEX Wikipedia
  - 653338 Wikipedia articles in early 2006
  - 50 entity ranking queries
  - Relevance assesments
- Our approach uses:
  - IR search
  - Link Analysis
  - Natural Language Processing (in the query)
  - Named Entity Recognition (in the query)

Outline

- Formal Model for Entity Ranking
- Application Scenarios
- Wikipedia Setting
- **Entity Ranking Algorithms**
- Experimental Results
- Demo
- Conclusions

Entity Ranking Algorithms

- INEX query
  - Keywords, Category
- Baseline
  - Search Keywords in the text and Category
- Link based
  - Links in Wikipedia are usually entities
  - Search Keywords also in anchor text of outLinks
- Synonyms and Related Words
  - Query extension: synonyms of nouns in the Keywords + Word Sense Disambiguation for the correct meaning

Entity Ranking Algorithms

- Core Characteristics
  - Clean the Keywords removing terms (and synonyms) appearing in Category
  - Keep only nouns and adjectives in Keywords
- Named Entities
  - Use only NE (i.e., organizations, locations, persons) from Keywords

Entity Ranking Algorithms

Title	Tom Hanks movies where he plays a leading role.
Category	Films
Synonyms	Tom ”Uncle Tom” Hanks ”Thomas J. Hanks” movies film flick ”motion picture” ”motion-picture show” ”moving picture” pic picture ”picture show” ”moving-picture show” where he plays a leading role
Related Words	**Synonyms** plus 50 additional concepts related mainly to motion pictures
Core Characteristics	Tom Hanks leading role
Named Entities	Tom Hanks

Experimental Results

- INEX XER 2007 benchmark
- MAP, P10 metrics

- Baseline

Nr	Method	P10
1	{text;Keywords}	0.19

- outLinks

Nr	Method	P10
2	{text;Keywords}; {outLinks;Keywords}	0.23
3	{text;Keywords}; {outLinks;CC(Keywords)}	0.26*
4	{text;Keywords}; {outLinks;NE(Keywords)}	0.24

Experimental Results

– Synonyms (SY) and Related Words (RW)

Nr	Method	P10
5	{text;Keywords U SY(Keywords)}	0.23
6	{text;Keywords U RW(Keywords)}	0.20

– Core Characteristics (CC) and Named Entities (NE)

Nr	Method	P10
7	{text;Keywords U CC(Keywords)}	0.23*
8	{text;Keywords U NE(Keywords)}	0.23*

Experimental Results

– Combinations

Nr	Method	P10
9	{text;Keywords U SY(Keywords) U RW(Keywords) U CC(Keywords) U NE(Keywords)}	0.28*
10	{text;Keywords U SY(Keywords) U RW(Keywords) U CC(Keywords) U NE(Keywords) }; {outLinks;CC(Keywords)}	0.29*

Demo

- American countries
- http://okkam.l3s.uni-hannover.de:8081/er08web/
- http://search.yahoo.com

Conclusions

- We presented a model for Entity Ranking
- The model can be applied to different scenarios
- We applied it to Wikipedia
- We defined algorithms for the Wikipedia context
- Results show that:
  - Combining Links,NLP,NER techniques we achieve 35% (MAP) and 53% (P10) improvement over normal search
  - Effectivness is overall low: young research area
- Next we will focus on searching the Web of Entities

Thanks