

# Entity Retrieval in the Knowledge Graph with Hierarchical Entity **Type** and **Content**

**Xinshi Lin**, Wai Lam and Kwun Ping Lai  
The Chinese University of Hong Kong

# Introduction

- **Entity**: broadly defined as a thing or concept that exists in the world or fiction
  - such as a person, a company or a color.
- **Entity Retrieval**: answering entity-targeted queries
  - e.g. 'give me all US food company stocks'
  - by returning a ranked list of entities from a structured knowledge base or knowledge graph.
    - e.g. McDonald's(NYSE:MCD), Burger King(NYSE:BKC), Kellogg's(NYSE:K) ...

Incandescent light bulb > Inventors

Thomas Edison



Hiram Maxim



Joseph Swan



← Entities

Who Invented the Light Bulb? - Live Science

<https://www.livescience.com/history>

Aug 16, 2017 - **Edison** and his team of researchers in Edison's laboratory in Menlo Park, N.J., tested more than 3,000 designs for bulbs between 1878 and 1880. In November 1879, **Edison** filed a patent for an electric lamp with a carbon filament.

Who really invented the light bulb? - Science Focus - BBC Focus ...

<https://www.sciencefocus.com/everyday-science>

US inventor **Thomas Edison** often gets all the credit, but was he really the ... In 1878, another British chemist, **Joseph Swan**, publicly demonstrated the first **light** ...

Videos



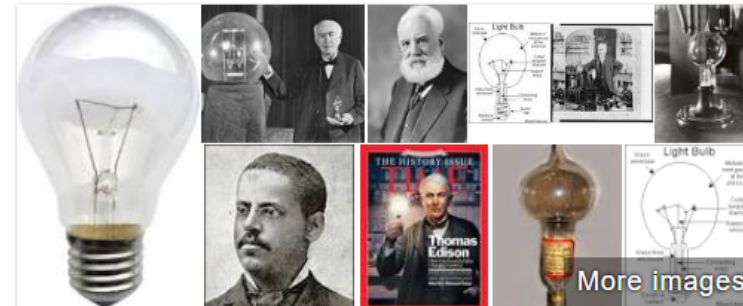
Ask History: Who Really Invented the Light Bulb? | History



Who invented the Lightbulb first?



Who Really Invented The Light Bulb - History of Everything



Incandescent light bulb

An incandescent light bulb, incandescent lamp or incandescent light globe is an electric light with a wire filament heated to such a high temperature that it glows with visible light. The filament is protected from oxidation with a glass or fused quartz bulb that is filled with inert gas or a vacuum. [Wikipedia](#)

People also search for

View 10+ more



Light-emitting diode



LED lamp



Lamp



Fluorescent lamp



Compact fluorescent

# Introduction – Entity Representation

- Entities are usually represented by subject-predicate-object(SPO) triples in the knowledge graph

<b>Ann_Dunham</b> - <b>&lt;rdfs:label&gt;</b> - <b>Ann Dunham</b>
<b>Ann_Dunham</b> - <b>&lt;dbo:abstract&gt;</b> - <b>Stanley Ann Dunham, the mother of Barack Obama, was an American anthropologist who ...</b>
<b>Ann_Dunham</b> - <b>&lt;dbo:birthPlace&gt;</b> - [ <b>&lt;Honolulu&gt;</b> , <b>&lt;Hawaii&gt;</b> ]
<b>Ann_Dunham</b> - <b>&lt;dbo:child&gt;</b> - <b>&lt;Barack Obama&gt;</b>
• • •

Corresponding SPO triples in DBpedia

## Knowledge base entry for ANN DUNHAM

```
<rdfs:label>:
  Ann Dunham


<dbo:abstract>:
  Stanley Ann Dunham, the mother of
  Barack Obama, was an American
  anthropologist who ...

<dbo:birthPlace>:
  [ <Honolulu>, <Hawaii> ]

<dbo:child>:
  <Barack_Obama>

<dbo:wikiPageWikiLink>:
  [ <United_States>,
    <Family_of_Barack_Obama>, ...]
```

**Ann Dunham**



**Born** Stanley Ann Dunham  
November 29, 1942  
Wichita, Kansas, U.S.

**Died** November 7, 1995 (aged 52)  
Honolulu, Hawaii, U.S.

**Cause of death** Complications from uterine cancer and ovarian cancer

**Education** University of Washington, Seattle  
University of Hawaii, Manoa (BA, MA, PhD)

**Spouse(s)** Barack Obama Sr. (m. 1961; div. 1964)  
Lolo Soetoro (m. 1965; div. 1980)

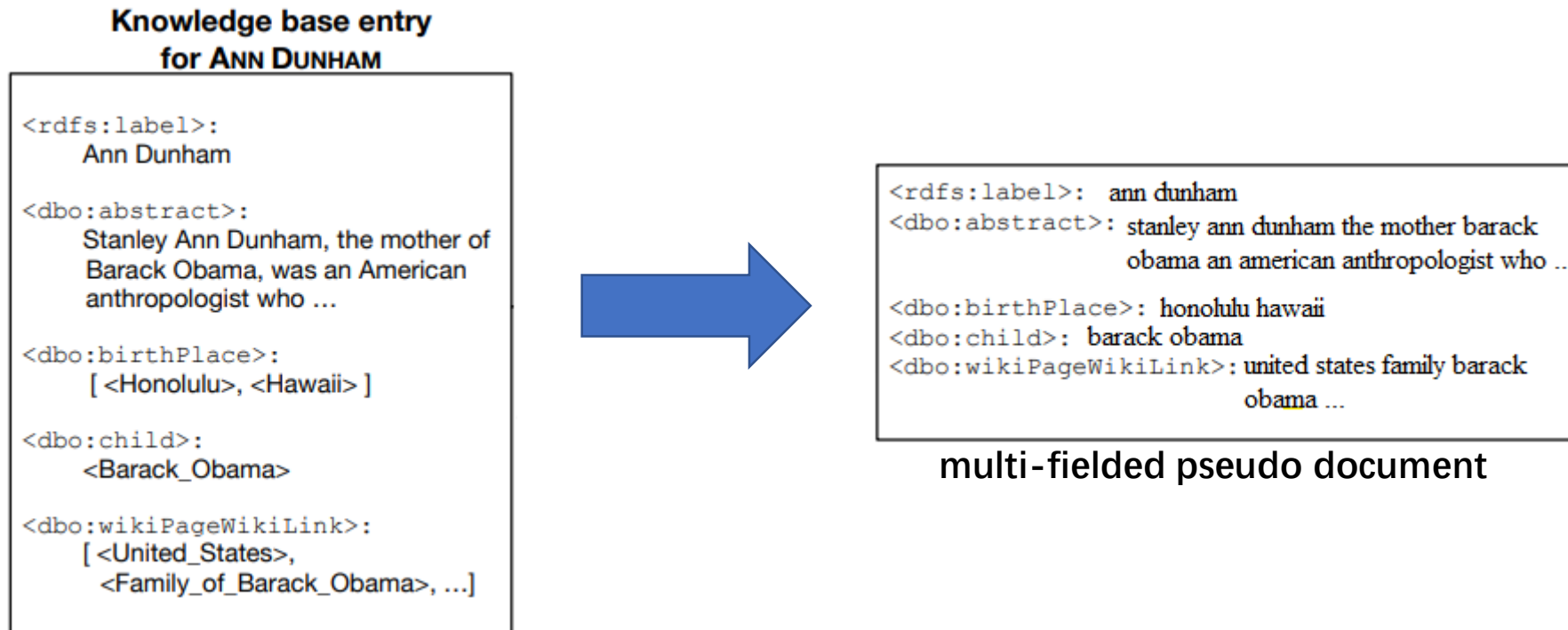
**Children** Barack Obama  
Maya Soetoro-Ng

**Parent(s)** Stanley Armour Dunham

Infobox in a corresponding Wikipedia Article

# Introduction – Previous Approaches

- Employ standard document retrieval methods
- by converting SPO triples into pseudo documents
  - e.g. “Steve\_Jobs-birthYear-1955” to “steve jobs birth year 1955”

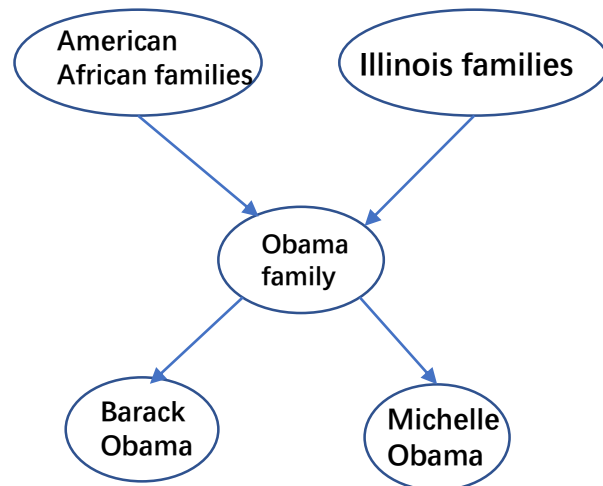


# Introduction

- However, there is more structure available than in standard document retrieval

## Entity Types in the **type taxonomy**

Categories: [Obama family](#) | [American anthropologists](#) | [American women](#) | [People from Honolulu](#) | [People from Mercer Island, Washington](#) | [People American people of Scottish descent](#) | [American people of Swiss descent](#) | [Deaths from uterine cancer](#) | [Mothers of Presidents of the United States](#)



## Entity Descriptions in the **knowledge source**

### Ann Dunham

From Wikipedia, the free encyclopedia

*Not to be confused with the British equestrian Anne Dunham.*

**Stanley Ann Dunham** (November 29, 1942 – November 7, 1995) was an American anthropologist who specialized in the economic anthropology and rural development of Indonesia.<sup>[1]</sup> She was the mother of Barack Obama, the 44th President of the United States.

Dunham was known as Stanley Ann Dunham through high school, then as **Ann Dunham**, **Ann Obama**, **Ann Soetoro**, **Ann Sutoro**, and finally after her second divorce as **Ann Dunham**.<sup>[2]</sup> Born in Wichita, Kansas, Dunham spent her childhood in California, Oklahoma, Texas and Kansas, her teenage years in Mercer Island, Washington, and most of her adult life in Hawaii and Indonesia.<sup>[3]</sup>

Dunham studied at the East–West Center and at the University of Hawaii at Manoa in Honolulu, where she attained a bachelor of arts degree in anthropology (1967),<sup>[4]</sup> and later received master of arts (1974) and PhD (1992) degrees, also in anthropology.<sup>[5]</sup> She also attended University of Washington at Seattle in 1961–1962. Interested in craftsmanship, weaving, and the role of women in cottage industries, Dunham's research focused on women's work on the island of Java and blacksmithing in Indonesia. To address the problem of poverty in rural villages, she created microcredit programs while working as a consultant for the United States Agency for International Development. Dunham was also employed by the Ford Foundation in Jakarta and she consulted with the Asian Development Bank in Gujranwala, Pakistan. Towards the latter part of her life, she worked with Bank Rakyat Indonesia, where she helped apply her research to the largest microfinance program in the world.<sup>[6]</sup>

After her son was elected President, interest renewed in Dunham's work: the University of Hawaii held a symposium about her research; an exhibition of Dunham's Indonesian batik textile collection toured the United States; and in December 2009, Duke University Press published *Surviving against the Odds: Village Industry in Indonesia*, a book based on Dunham's original 1992 dissertation. Janny Scott, an author and former *New York Times* reporter, published a biography about Ann Dunham's life titled *A Singular Woman* in 2011. Posthumous interest has also led to the creation of The Ann Dunham Soetoro Endowment in the Anthropology Department at the University of Hawaii at Manoa, as well as the Ann Dunham Soetoro Graduate Fellowships, intended to fund students associated with the East–West Center (EWC) in Honolulu, Hawaii.<sup>[6]</sup>

In an interview, Barack Obama referred to his mother as "the dominant figure in my formative years ... The values she taught me continue to be my touchstone when it comes to how I go about the world of politics."<sup>[7]</sup>

#### Contents [hide]

- 1 Early life
- 2 Family life and marriages
  - 2.1 First marriage
  - 2.2 Second marriage
- 3 Professional life
- 4 Illness and death
- 5 Posthumous interest
- 6 Personal beliefs

Ann Dunham	
<span></span>	
Born	<span>Stanley Ann Dunham</span> November 29, 1942 <div>Wichita, Kansas, U.S.</div>
Died	November 7, 1995 (aged 52) <div>Honolulu, Hawaii, U.S.</div>
Cause of death	Complications from uterine cancer and ovarian cancer
Education	University of Washington, Seattle <div>University of Hawaii, Manoa (BA, MA, PhD)</div>
Spouse(s)	<span>Barack Obama Sr.<span> </span>(m.<span> </span>1961<span>;</span><span> </span>div.<span> </span>1964)</span> <span>Lolo Soetoro<span> </span>(m.<span> </span>1965<span>;</span><span> </span>div.<span> </span>1980)</span>
Children	<span>Barack Obama</span> <span> </span> <span>Maya Soetoro-Ng</span>
Parent(s)	<span>Stanley Armour Dunham</span>

# Our proposal

- Improve entity retrieval by incorporating
  - hierarchical entity type information
  - entity descriptions.

# Model Description

- Given a query  $Q$
- A candidate entity  $E$
- $D = \{D_f\}$ : multi-fielded pseudo documents of  $E$ .

The overall scoring function

$$g(Q, D, E) = \lambda_W \tilde{g}(Q, D, E) + (1 - \lambda_W) \tilde{h}(Q, E)$$

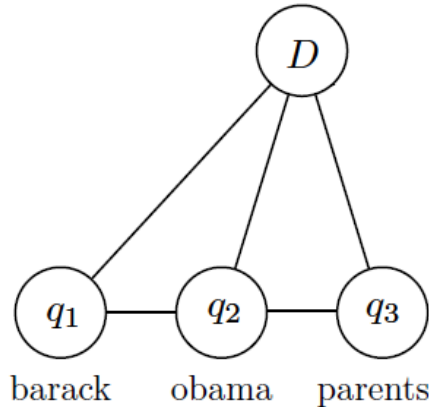
$\tilde{g}(Q, D, E)$ : scoring function incorporating entity type information

$\tilde{h}(Q, E)$ : scoring function incorporating entity descriptions

$\lambda_W$ : model parameter



# Model Description - Incorporate entity type information



**Markov random field** based framework

Example: query='barack Obama parents'

**unigrams**={'barack','Obama','parents'}

**bigrams**={'barack Obama', 'Obama parents'}

$$\tilde{g}(Q, D, E) = \max_{c \in \text{types}(E)} \max_{p \in \text{Paths}(c)} \left\{ \lambda_T \sum_{q_i \in Q} \tilde{f}_T(q_i, D, p) + \right.$$

feature function for unigrams

$$\lambda_O \sum_{q_i, q_{i+1} \in Q} \tilde{f}_O(q_i, q_{i+1}, D, p) +$$

feature function for ordered  
bigram occurrences

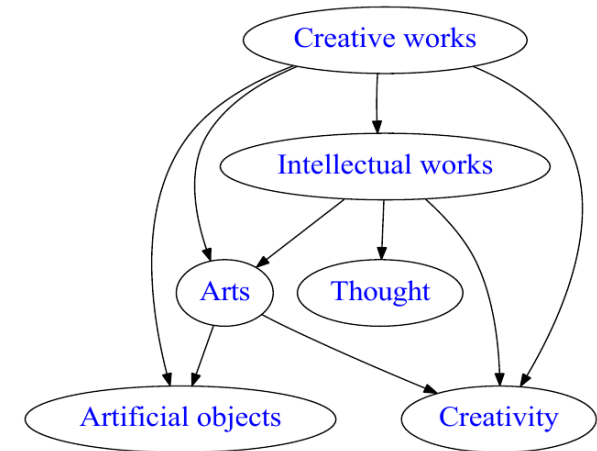
$$\lambda_U \sum_{q_i, q_{i+1} \in Q} \tilde{f}_U(q_i, q_{i+1}, D, p) \}$$

feature function for unordered  
bigrams occurrences

# Model Description - Incorporate entity type information

- Consider path as a variable
- Enumerate each path  $p$  in the type taxonomy
  - starts from an entity type  $c$

$$\begin{aligned}\tilde{g}(Q, D, E) = & \max_{c \in \text{types}(E)} \max_{p \in \text{Paths}(c)} \{ \lambda_T \sum_{q_i \in Q} \tilde{f}_T(q_i, D, p) + \\ & \lambda_O \sum_{q_i, q_{i+1} \in Q} \tilde{f}_O(q_i, q_{i+1}, D, p) + \\ & \lambda_U \sum_{q_i, q_{i+1} \in Q} \tilde{f}_U(q_i, q_{i+1}, D, p) \}\end{aligned}$$



# Model Description - Incorporate entity type information

- Dirichlet prior smoothed feature function

$$\tilde{f}_{\{T,O,U\}}(W, D, p) = \log \sum_f w_f^{\{T,O,U\}} \frac{tf_{W,D_f} + N_t}{|D_f| + D_t}$$

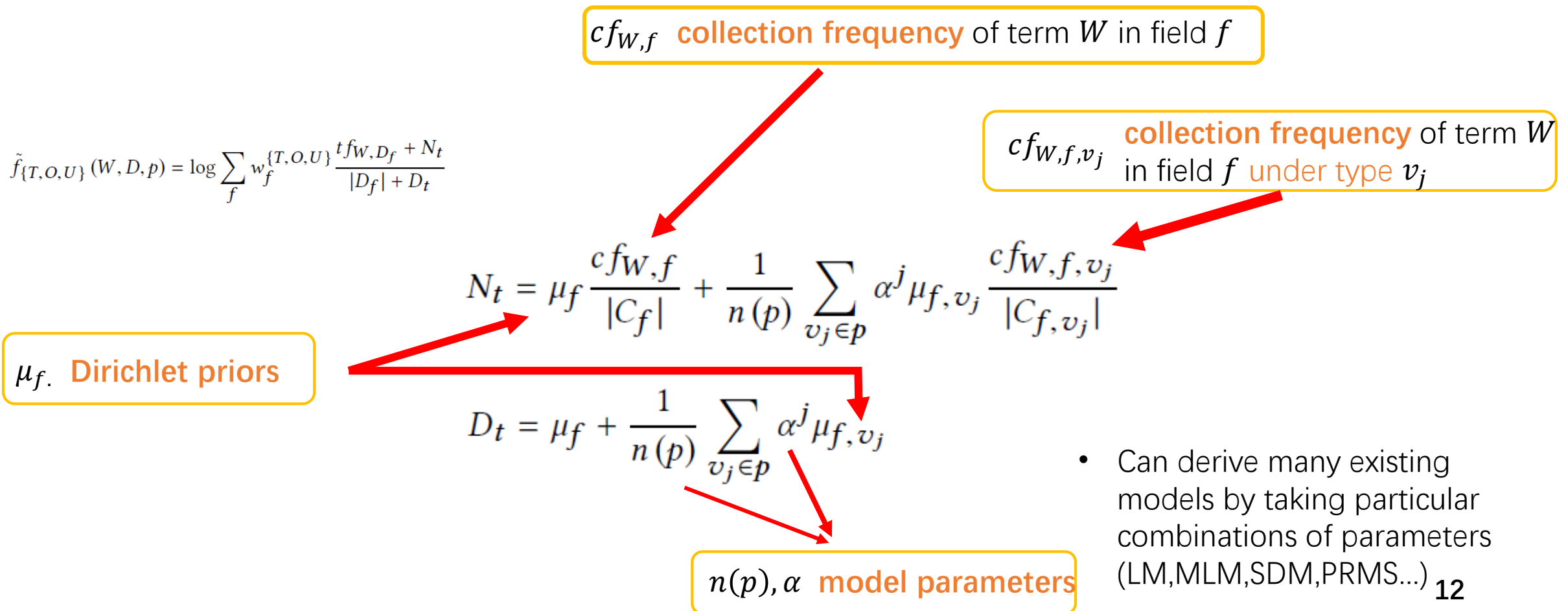
$tf_{W,D_f}$ : term frequency of  $W$  in the pseudo document of  $f$

$W$  generalizes to unigrams and bigrams

$N_t, D_t$ : path-aware smoothing components

# Model Description - Incorporate entity type information

- Specifically, we add a **path-aware smoothing component** in the feature function



# Model Description – Time Complexity Analysis

$$T = O(|types(E)| \cdot |paths(c)| \cdot |p| \cdot |Q| \cdot |F|)$$

$|types(E)|$  number of entity types considered  
 $|paths(c)|$  number of paths explored in the type taxonomy  
 $|p|$  average length of a path  
 $|Q|$  average length of a query  
 $|F|$  number of fields

- Given a query and a candidate entity
- assuming the operation of computing term frequencies and collection frequencies takes constant time
- Path exploration implemented in a recursive way

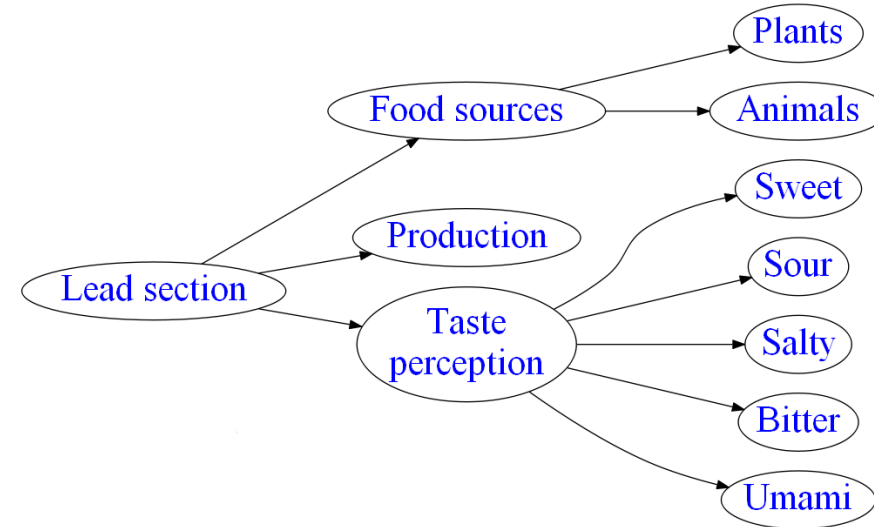
# Model Description - Incorporating entity descriptions

- Similar formulation
- Enumerate each path  $p$  from lead section to a leaf section

$$\tilde{h}(Q, E) = \max_{p \in T(E)} \left\{ \lambda_T \sum_{q_i \in Q} \tilde{h}_T(q_i, E, p) + \lambda_O \sum_{q_i, q_{i+1} \in Q} \tilde{h}_O(q_i, q_{i+1}, E, p) + \lambda_U \sum_{q_i, q_{i+1} \in Q} \tilde{h}_U(q_i, q_{i+1}, E, p) \right\}$$

$$\tilde{h}_{\{T, O, U\}}(W, E, p) = \log \frac{\sum_{s_j \in p} \beta^j \cdot t f_{W, s_j} + \mu_d \frac{cf_{W, C_d}}{|C_d|}}{\sum_{s_j \in p} \beta^j \cdot |s_j| + \mu_d}$$

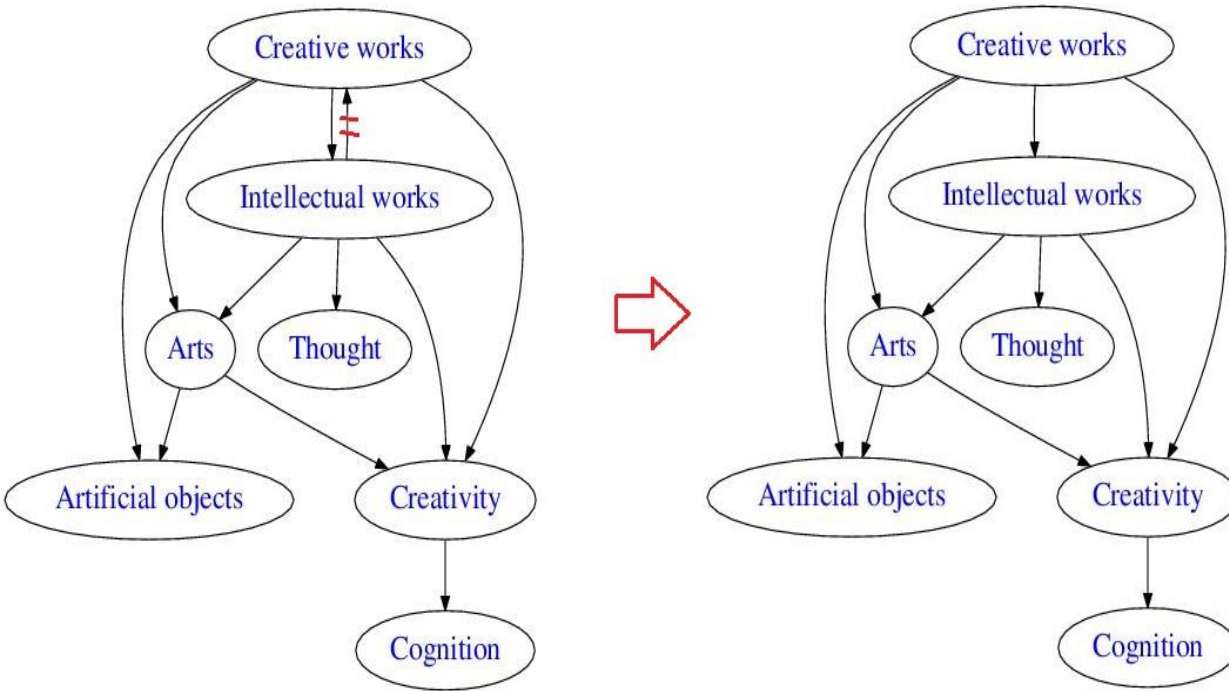
$s_j$ : the  $j$ -th section in the path



# Experiment

- Knowledge graph: DBpedia 2015-10
- Knowledge source: Wikipedia 2015-10
- Type taxonomy: Wikipedia Category System
- Test collection: DBpedia-Entity v2
  - Four benchmark datasets with 467 queries in total
    - INEX-LD: keyword based queries
    - SemSearch\_ES: named entity targeted queries
    - ListSearch: queries that seek a particular list of entities
    - QALD2: natural language questions
  - Metric: NDCG@10, MAP@100
  - “DBpedia-Entity v2: A Test Collection for Entity Search”, Hasibi et.al, SIGIR 2017

# Wikipedia Category Processing



## Step 1 Construct Wikipedia Category Graph

Each node represents a Wikipedia category and connects to its parent categories.

## Step 2 Remove Strongly Connected Components in the Graph

Divide the graph into SCCs. A SCC is then reduced by removing common edges shared by intersected elementary cycles or one edge of a sole circle.



# Experiment

## Baselines

- **BM25F**: the BM25 Model with extension to multiple weighted fields  
Stephen Robertson, Hugo Zaragoza, et al. (2009)
- **PRMS**: the Probabilistic Retrieval Model for Semi structured Data  
Jinyoung Kim, Xiaobing Xue, and W Bruce Croft. (2009)
- **LM**: standard language modelling  
Zhai Chengxiang, and John Lafferty. (2004)
- **MLM**: Mixture of Language models  
Paul Ogilvie and Jamie Callan. (2003)
- **SDM**: Sequential Dependence Model  
Donald Metzler and W Bruce Croft. (2005)
- **FSDM**: Fielded Sequential Dependence Model  
Nikita Zhiltsov, Alexander Kotov, and Fedor Nikolaev. (2015)  
Faegheh Hasibi, Krisztian Balog, and Svein Erik Bratsberg. (2016)

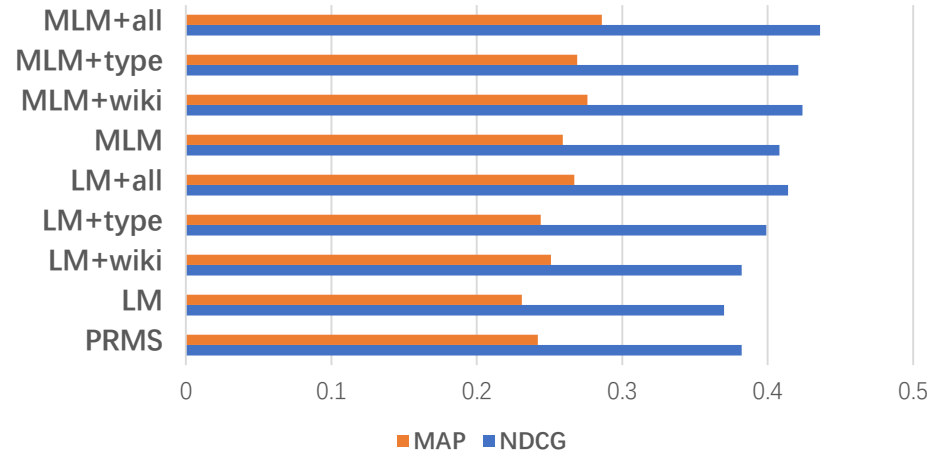
**Table 1: Results of our approach with the Wikipedia category information and the Wikipedia articles. The symbol \* denotes that the improvements over the corresponding existing model are statistically significant ( $p < 0.05$ ).**

Model	INEX-LD		SemSearch ES		ListSearch		QALD2	
	NDCG	MAP	NDCG	MAP	NDCG	MAP	NDCG	MAP
BM25F	.413	.253	.624	.488	.380	.279	.337	.248
PRMS	.382	.242	.663	.529	.378	.267	.348	.247
LM	.370	.231	.574	.482	.389	.284	.348	.243
LM+wiki	.392*	.251*	.598*	.480	.397	.297*	.353	.256*
LM+type	.399*	.244*	.552	.461	.398*	.301*	.352	.258*
LM+all	.414*	.267*	.593*	.485	.410*	.323*	.369*	.276*
MLM	.408	.259	.676	.547	.372	.278	.358	.267
MLM+wiki	.424*	.276*	.681	.552	.380	.289*	.371*	.277*
MLM+type	.421*	.269*	.672	.543	.374	.283	.359	.271
MLM+all	.436*	.286*	.683*	.552*	.388*	.298*	.374*	.284*
SDM	.373	.233	.604	.510	.394	.288	.354	.251
SDM+wiki	.395*	.256*	.618*	.502	.408*	.304*	.360	.264*
SDM+type	.383*	.239*	.607	.517	.401*	.296*	.360	.259
SDM+all	.417*	.266*	.616*	.515	.423*	.320*	.376*	.276*
FSDM	.394	.251	.672	.547	.406	.289	.360	.260
FSDM+wiki	.412*	.270*	.671	.543	.407	.304*	.388*	.276*
FSDM+type	.402*	.247	.662	.539	.402	.282	.372*	.266
FSDM+all	.412*	.269*	.668	.534	.407	.290	.374*	.272*

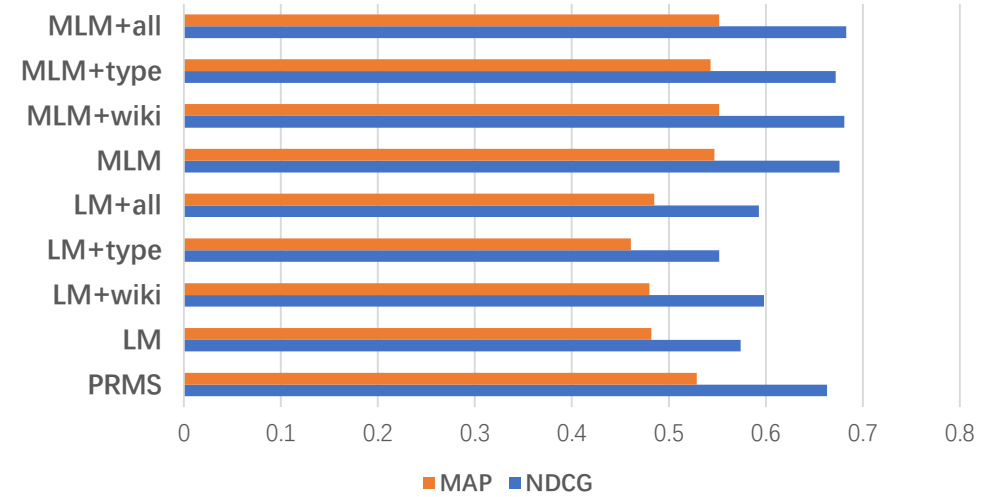
# Experiment

## Baselines – LM based ones

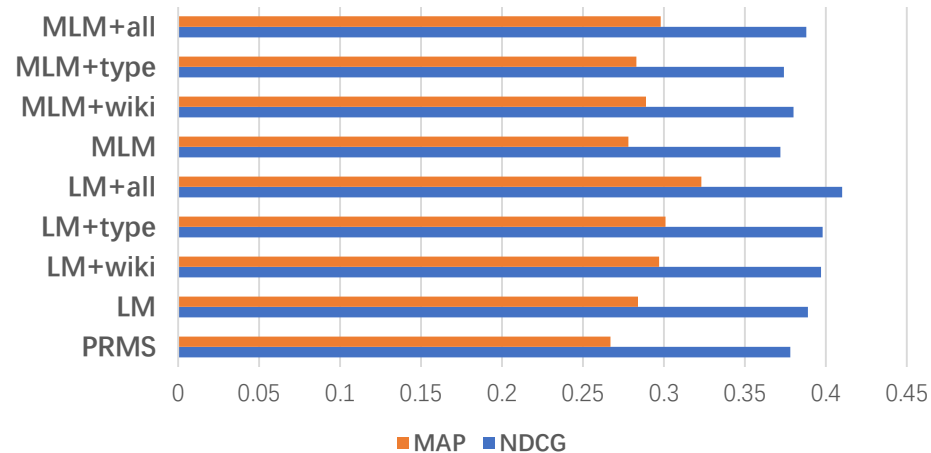
INEX-LD



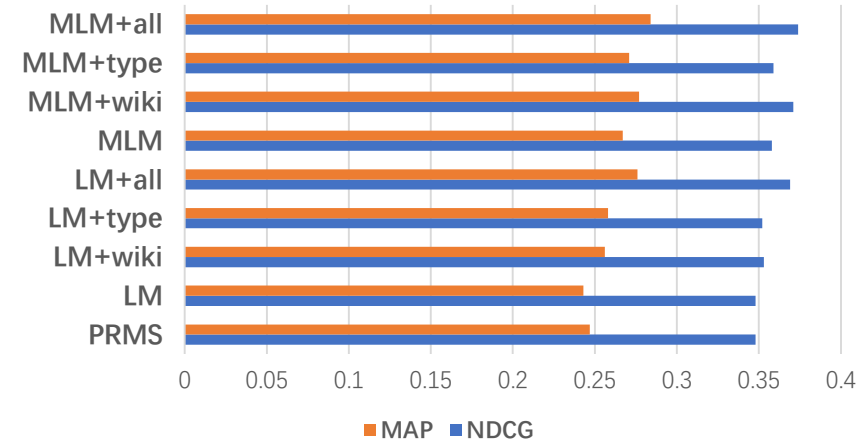
SemSearch-ES



ListSearch



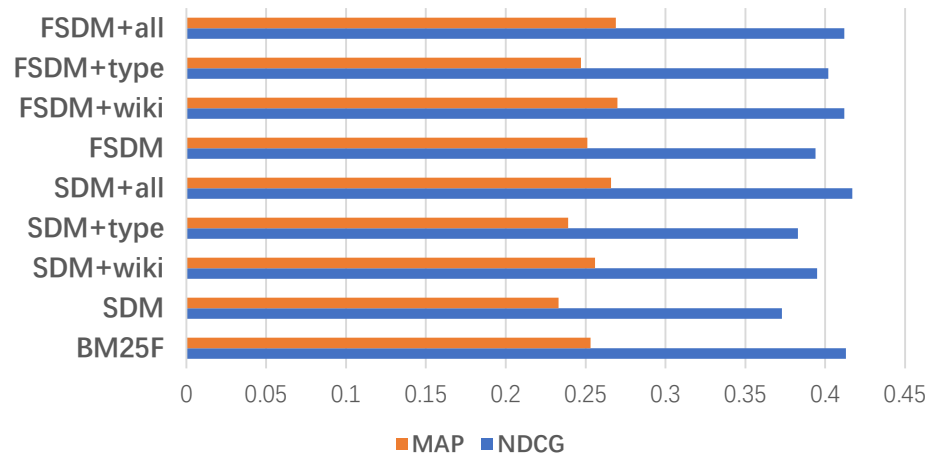
QALD2



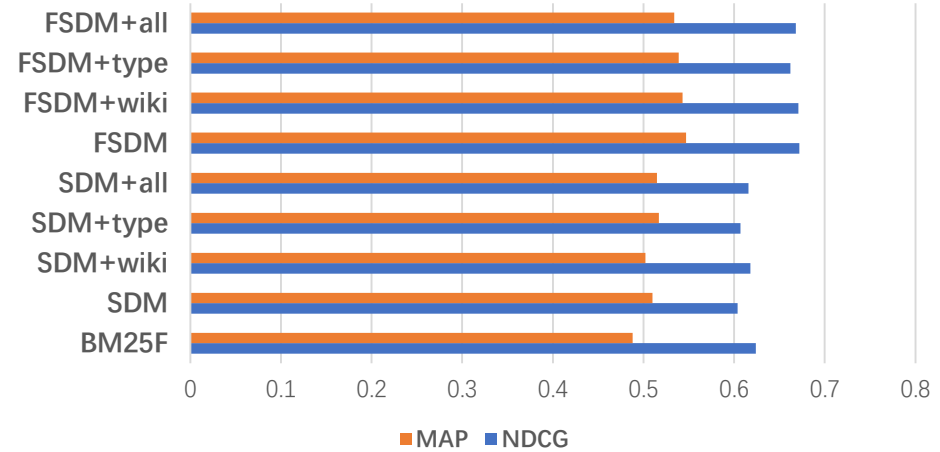
# Experiment

## Baselines – MRF based ones and others

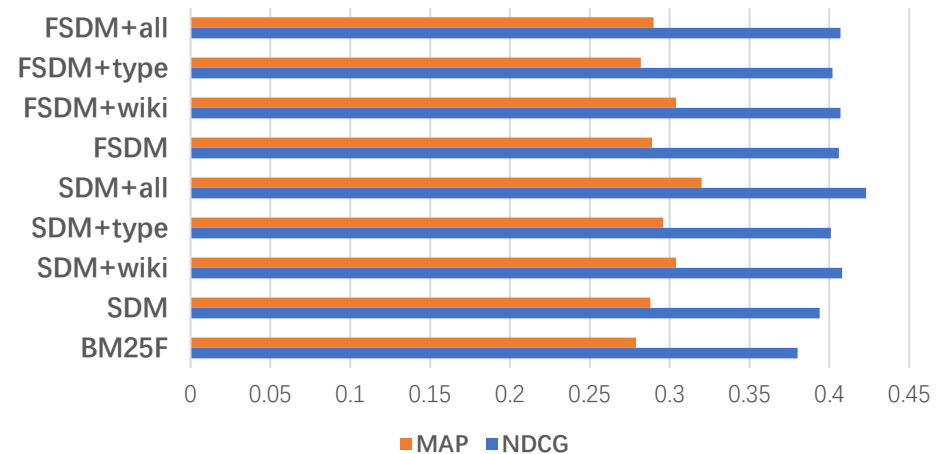
INEX-LD



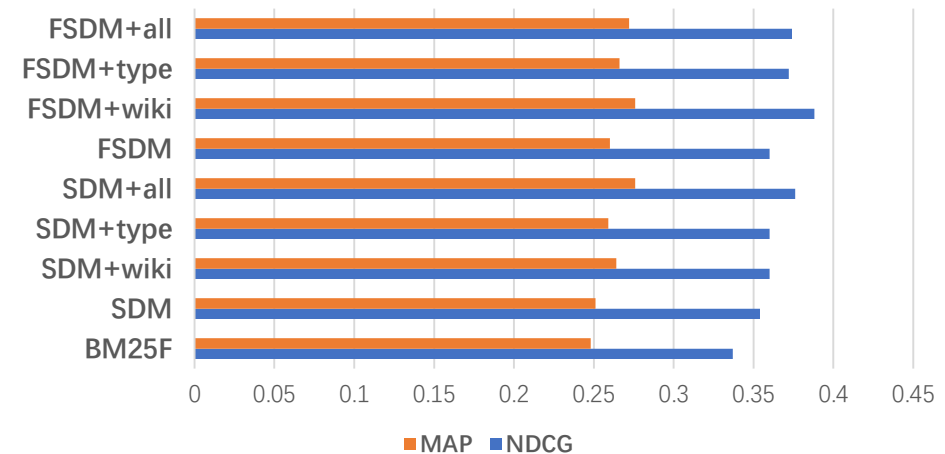
SemSearch-ES



ListSearch

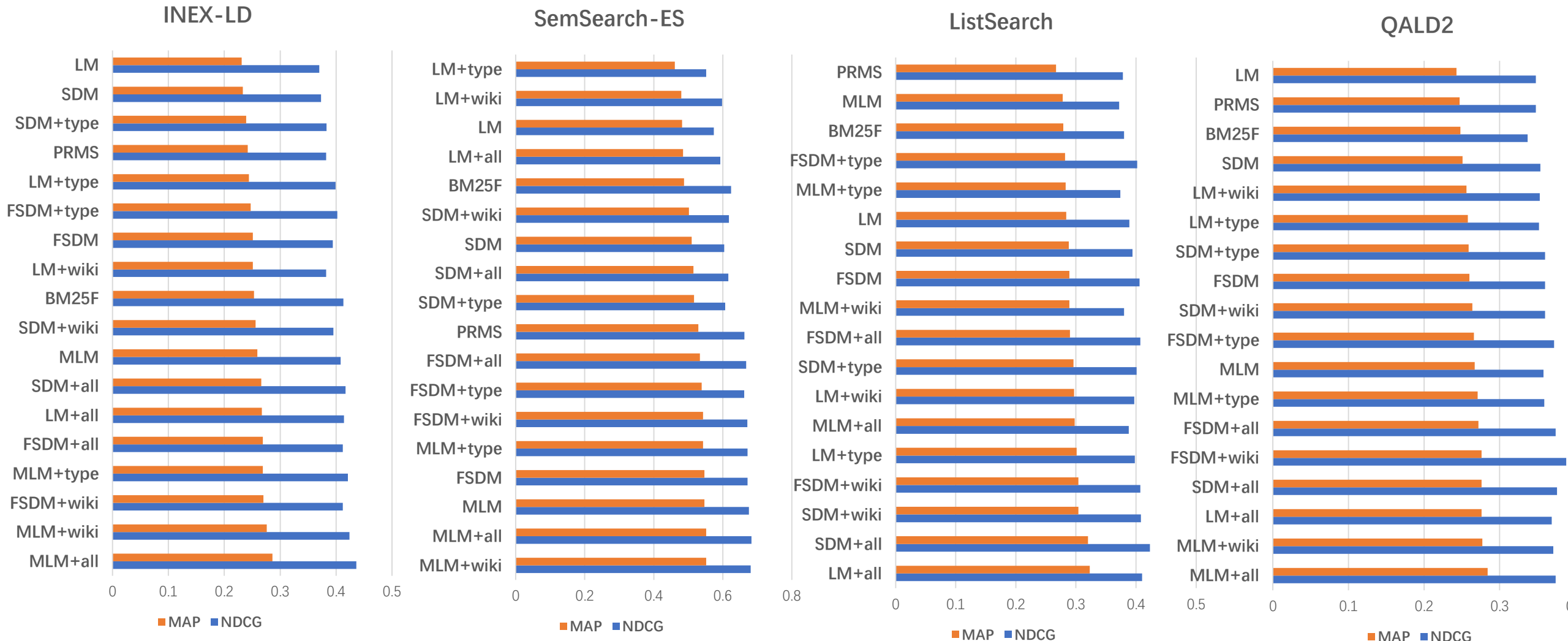


QALD2



# Experiment

## Dataset specific results



# Experiment

## Analysis

- Incorporating both entity documents and type information brings the largest improvements than with either of them.
- For all dataset except SemSearch ES, the entity descriptions and type information contribute roughly equally
- On SemSearch ES, most improvements comes from exploiting the entity descriptions
- Due to different query characteristics/intents in each dataset.

**Table 1: Results of our approach with the Wikipedia category information and the Wikipedia articles. The symbol \* denotes that the improvements over the corresponding existing model are statistically significant ( $p < 0.05$ ).**

Model	INEX-LD		SemSearch ES		ListSearch		QALD2	
	NDCG	MAP	NDCG	MAP	NDCG	MAP	NDCG	MAP
BM25F	.413	.253	.624	.488	.380	.279	.337	.248
PRMS	.382	.242	.663	.529	.378	.267	.348	.247
LM	.370	.231	.574	.482	.389	.284	.348	.243
LM+wiki	.392*	.251*	.598*	.480	.397	.297*	.353	.256*
LM+type	.399*	.244*	.552	.461	.398*	.301*	.352	.258*
LM+all	.414*	.267*	.593*	.485	.410*	.323*	.369*	.276*
MLM	.408	.259	.676	.547	.372	.278	.358	.267
MLM+wiki	.424*	.276*	.681	.552	.380	.289*	.371*	.277*
MLM+type	.421*	.269*	.672	.543	.374	.283	.359	.271
MLM+all	.436*	.286*	.683*	.552*	.388*	.298*	.374*	.284*
SDM	.373	.233	.604	.510	.394	.288	.354	.251
SDM+wiki	.395*	.256*	.618*	.502	.408*	.304*	.360	.264*
SDM+type	.383*	.239*	.607	.517	.401*	.296*	.360	.259
SDM+all	.417*	.266*	.616*	.515	.423*	.320*	.376*	.276*
FSDM	.394	.251	.672	.547	.406	.289	.360	.260
FSDM+wiki	.412*	.270*	.671	.543	.407	.304*	.388*	.276*
FSDM+type	.402*	.247	.662	.539	.402	.282	.372*	.266
FSDM+all	.412*	.269*	.668	.534	.407	.290	.374*	.272*

# Summary

- Propose a Markov random field based framework incorporating both hierarchical entity type information and entity descriptions
  - Add a path aware smoothing component in the feature functions
- Propose a graph-based approach to process Wikipedia category system
- Comparing to existing type-aware entity retrieval models, our model does not require explicit inference of query type
- Future work includes investigation of incorporating more knowledge graph structures and query characteristics/intents

Thank you

SIGIR Student Travel Grant