# Hidden Markov Model for Term Weighting in Verbose Queries

Xueliang Yan, Guanglai Gao, Xiangdong Su, Hongxi Wei,
Xueliang Zhang, Qianqian Lu

**College of Computer Science**

**Inner Mongolia University**

**Huhhot 010021, China**

# Outline

- Introduction
- Basic Idea
- Experiments
- Conclusion

# Introduction

- Current search engines perform well with keyword queries

  time, conference, CLEF2012

- but are not, in general, effective for verbose queries.

  ‘Can you tell me the exact time that the conference of CLEF2012 will be hold. ……………………………………’

→ The main reason for this is that most retrieval models treat all the terms in the query as equally important (an assumption that often does not hold)

# Basic Idea

- Term POS ←→ Term Weight
  - Noun → important;
  - Prep → non-important
- Term organization ←→ Term Weight

NN+IN+NN:

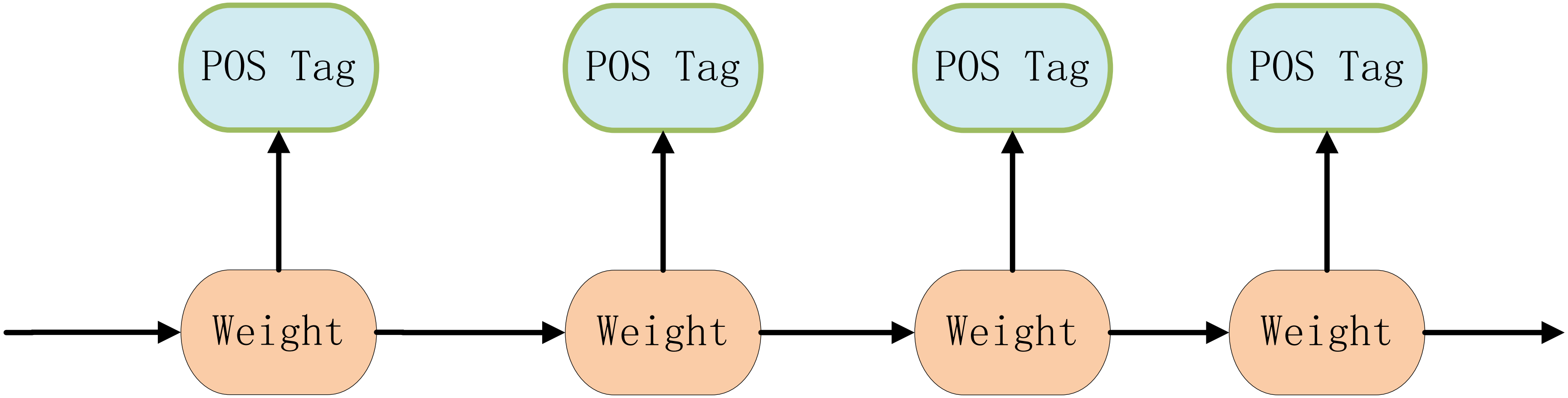
- description of nature;
- quality of life;
- extinction of wildlife;
- use of estrogen
- …

NN+NN+IN:

- air pollution in
- owl episode in
- life style of
- Tobacco industry for
- …

- Capture the above relationships

# Basic Idea



Max $P(\text{Weight}_1, \text{Weight}_2, \ldots, \text{Weight}_n | \text{POS}_1, \text{POS}_2, \ldots, \text{POS}_n)$

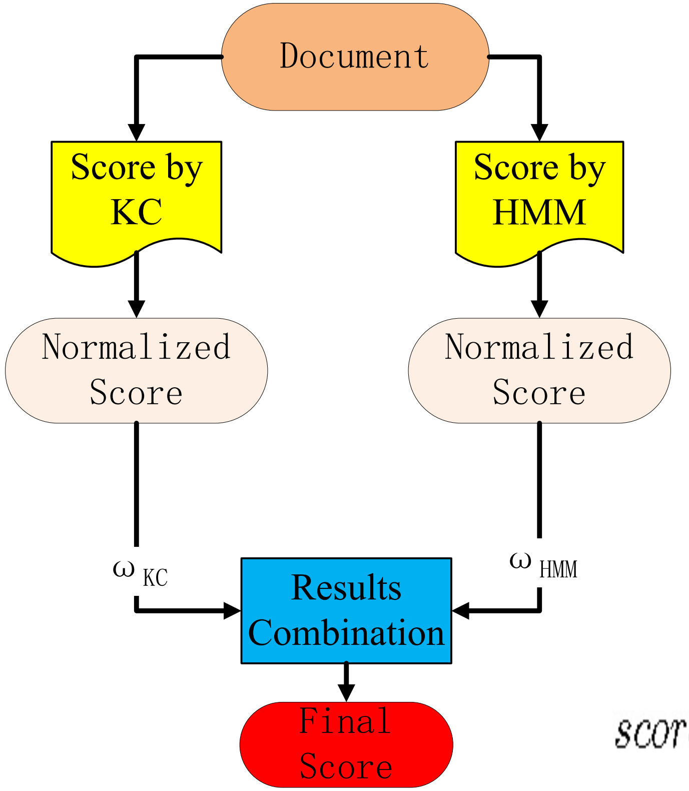
# Experiments

- TREC Robust04 track
- 250 topics
- Indri
- Indri Query Language

# Experiment Results

| | TopicSet_1 | | TopicSet_2 | | TopicSet_3 | | TopicSet_4 | | TopicSet_5 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | MAP | P@5 | MAP | P@5 | MAP | P@5 | MAP | P@5 | MAP | P@5 |
| Query Likelihood | 0.184 | 0.348 | 0.157 | 0.4 | 0.215 | 0.436 | 0.326 | 0.567 | 0.279 | 0.5 |
| OKAPI | 0.188 | 0.348 | 0.165 | 0.425 | 0.221 | 0.432 | 0.321 | 0.551 | 0.279 | 0.508 |
| KC | 0.212 | 0.356 | 0.196 | 0.468 | 0.226 | 0.44 | 0.343 | 0.552 | 0.308 | 0.571 |
| HMM | 0.213 | 0.368 | 0.189 | 0.468 | 0.224 | 0.444 | 0.335 | 0.564 | 0.291 | 0.514 |

# Results Combination



$$score_{\mathrm{K+H}}(d) = w_{\mathrm{KC}} * norm\_score_{\mathrm{KC}}(d) + w_{\mathrm{HMM}} * norm\_score_{\mathrm{HMM}}(d)$$

| | TopicSet_1 | | TopicSet_2 | | TopicSet_3 | | TopicSet_4 | | TopicSet_5 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | MAP | P@5 | MAP | P@5 | MAP | P@5 | MAP | P@5 | MAP | P@5 |
| KC | 0.212 | 0.356 | 0.196 | 0.468 | 0.226 | 0.44 | 0.343 | 0.552 | 0.308 | 0.571 |
| HMM | 0.213 | 0.368 | 0.189 | 0.468 | 0.224 | 0.444 | 0.335 | 0.564 | 0.291 | 0.514 |
| KC+HMM | 0.219 | 0.368 | 0.202 | 0.476 | 0.23 | 0.448 | 0.35 | 0.576 | 0.309 | 0.563 |

# Conclusion

- Both POS and the Organization of term have relationship with the importance of term
- HMM can capture such information to determine term weight
- There is potential to be combined with other models that used different information

# Future Work

- Not linear, more complex, like a tree
- Other combination method

# Thanks!

## Comments & Questions?

College of Computer Science
Inner Mongolia University
Huhhot 010021, China
Email: csggl@imu.edu.cn
Tel:+86-471-4992341