A Few Things to Know about Machine Learning for Web Search

Hang Li
Noah’s Ark Lab
Huawei Technologies

AIRS 2012
Tianjin, China
Dec. 19, 2012
Talk Outline

• My projects at MSRA
• Some conclusions from our research on web search
My Past Projects at MSRA

• Text Mining (2002-2005)
  – Development of SQL Server 2005 Text Mining

• Enterprise Search (2003-2012)

• Web Search (2005-2012)
  – Development of Live Search 2008, Bing 2009
Research on Machine Learning for Web Search

• Learning to Rank
  – **Tie-Yan Liu, Jun Xu, Tao Qin, etc**
  – Letor dataset [Liu+ 07], ListNet[Cao+ 07], ListMLE[Xia+ 09], AdaRank[Xu+07], IR SVM [Cao+ 06]

• Importance Ranking
  – **Tie-Yan Liu, Bin Gao, etc**
  – BrowseRank [Liu+ 08]

• Semantic Matching (Relevance)
  – **Gu Xu, Jun Xu, Jingfang Xu, etc**
  – CRF [Guo+ 08], NERQ [Guo+ 09], LogLinear [Wang+ 11], RLSI [Wang+ 11], RMLS[Wu+ 12], SRK [Bu+ 12]

• Search Log Mining
  – **Daxin Jiang, Yunhua Hu, etc**
  – Context-aware Search [Cao+ 08] [Cao+ 09][Xiang +11], Intent Mining [Hu+ 12]
Research on Machine Learning for Web Search (cont’)

• We tried to address the fundamental ‘computer science problems’, i.e., to develop fundamental models (algorithms)

• Performance can be further improved by adding engineering efforts
Some Conclusions from Our Research

- Machine learning based ranking and rule-based ranking both have pros and cons
- State of the art learning to rank algorithms
- More features better performance
- No signal for relevance is enough
- Matching (feature) is more important than ranking (model)
- Matching can be performed at multiple levels
- Click data is useful
- Browse data is useful
- Flexibility is key for handling queries
- List of useful features in ranking
- Spelling errors in query can be corrected first
Beyond Search

• Other applications have similar problems
  – Online advertisement
  – Question answering
  – Recommender system
  – … …

• Techniques can be applied to the applications as well
Machine Learning based Ranking vs Rule based Ranking

• Two types of signals
  – Relevance (matching)
  – Importance
  – The higher the scores are, the better relevance is

• Simplest model
  – Linear combination

• Make it possible for rule-based approach

• Precise tuning needs either learning-based approach (learning to rank) or rule-based approach
# Machine Learning based Ranking vs Rule based Ranking

<table>
<thead>
<tr>
<th></th>
<th>Learning based</th>
<th>Rule based</th>
</tr>
</thead>
<tbody>
<tr>
<td>Update of model</td>
<td>Easy</td>
<td>Hard</td>
</tr>
<tr>
<td>Fine tuning</td>
<td>Hard to control</td>
<td>Easy to control</td>
</tr>
<tr>
<td>Creation of model</td>
<td>Optimized for average cases</td>
<td>Can be optimized to avoid worst cases</td>
</tr>
<tr>
<td>Creation of training data</td>
<td>Necessary</td>
<td>Not necessary</td>
</tr>
</tbody>
</table>
State of the Art Learning to Rank Algorithms

- LambdaMart
- LambdaRank
- ListNet
- AdaRank
- Rank SVM
- IR SVM
- RankNet
- RankBoost
- LambdaMark performed the best in Yahoo Competition, etc.
- The differences among the above rankers are small
More Features Better Performance

• The more features used in ranker (ranking model), usually the better performance
• Even ‘redundant’ features (e.g., BM25 and tf-idf)
• In terms of NDCG and the Cranefield evaluation
No Signal (Feature) is Enough

- Not possible to just use one type of signal
- Power law distribution (long tail)
- Head is easy, but tail is hard
- Representing signals at
  - Multiple fields: title, anchor, url, click
Matching (Feature) vs Ranking (Model)

• In traditional IR:
  – Ranking = matching
    \[ f(q,d) = f_{BM25}(q,d) \text{ or } f(q,d) = P_{LMIR}(d \mid q) \]

• Web search:
  – Ranking and matching become separated
  – Learning to rank becomes state-of-the-art
    \[ f(q,d) = f_{BM25}(q,d) + g_{PageRank}(d) + \cdots \]
  – Matching = feature learning for ranking

• Learning to Match
Same Search Intent Different Query Representations
Example = “Distance between Sun and Earth”

• "how far" earth sun
• "how far" sun
• "how far" sun earth
• average distance earth sun
• average distance from earth to sun
• average distance from the earth to the sun
• distance between earth & sun
• distance between earth and sun
• distance between earth and the sun
• distance from earth to the sun
• distance from sun to earth
• distance from sun to the earth
• distance from the earth to the sun
• distance from the sun to earth
• distance from the sun to the earth
• distance of earth from sun
• distance between earth sun
• how far away is the sun from earth
• how far away is the sun from the earth
• how far earth from sun
• how far earth from the sun
• how far earth is from the sun
• how far from earth to sun
• how far from earth is the sun
• how far from earth to sun
• how far from the earth to the sun
• distance between sun and earth
Matching at Multiple Levels

Match between structures of query & document title

how far is sun from earth → ... distance between sun and earth

Match between topics of query & document

Microsoft Office → ... Microsoft ... PowerPoint, Word, Excel...

Match between word senses in query & document

utube → youtube     NY → New York

Match between phrases in query & document

hot dog → hot dog

Match between terms in query & document

NY → NY     youtube → youtube
Click Data

• Queries associated with page in click data can be viewed as metadata of page
• Useful streams (fields): title, anchor, url, click, and body
• Web search technologies
  – First generation: traditional IR
  – Second generation: anchor text, PageRank
  – Third generation: click data, learning to rank, etc
Browse Data

• PageRank is not as powerful as people may expect
• Number of visits is a good strong for page importance
• BrowseRank (continuous time Markov process)
Flexibility Is Key for Handling Queries

• Four types of queries
  – Noun phrases
  – Multiple noun phrases
  – Titles of books, songs, etc
  – Natural language questions (about 1%)

• Needs to handle variants of expressions (cf., distance between sun and earth)

• String Re-writing Kernel (Bu 2012) for tackling flexibility of queries
List of Useful Features

• Features can be defined in multiple fields
  – Title
  – Anchor
  – URL
  – Click
  – Body

• Useful features
  – BM25
  – N-gram BM25
  – Exact match
  – Translation between queries and titles
  – Topic model
  – Latent matching model
  – PageRank
  – BrowseRank
Spelling Error Correction

• 10-15 English queries contain spelling errors
• Formalized as string transformation problem
• CRF [Guo et al 08]
• Spelling error correction should be done only when confident
  – Eg. ‘mlss singapore’ = ‘miss singapore’ or ‘machine learning summer school singapore’
• Spelling error correction does not depend on documents
• Other query re-writing depends on documents
  – E.g, ‘seattle best hotel’ vs ‘seattle best hotels’
  – Eg., ‘arms reduction’ vs ‘arm reduction’
Some Conclusions from Our Research

• Machine learning based ranking and rule-based ranking both have pros and cons
• State of the art learning to rank algorithms
• More features better performance
• No signal for relevance is enough
• Matching (feature) is more important than ranking (model)
• Matching can be performed at multiple levels
• Click data is useful
• Browse data is useful
• Flexibility is key for handling queries
• List of useful features in ranking
• Spelling errors in query can be corrected first
References

• Wei Wu, Zhengdong Lv, Hang Li, Regularized Mapping to Latent Structures and Its Application to Web Search, under review.


• Hang Li, A Short Introduction to Learning to Rank, IEICE Transactions on Information and Systems, E94-D(10), 2011.


References


• Huanhuan Cao, Daxin Jiang, Jian Pei, Qi He, Zhen Liao, Enhong Chen, Hang Li. Context-Aware Query Suggestion by Mining Click-Through and Session Data, In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'08), pages 875-883, 2008. (SIGKDD’08 Best Application Paper Award).


References

Thank You!

Contact: hangli.hl@huawei.com