Learning Maximal Marginal Relevance Model via Directly Optimizing Diversity Evaluation Measures

Long Xia, Jun Xu, Yanyan Lan, Jiafeng Guo, Xueqi Cheng

Key Laboratory of Network Data Science and Technology
Institute of Computing Technology
Chinese Academy of Sciences
Outline

• Background
• Related work
• Our approach
• Experiments
• Summary
Problem of diversity

Jaguar
https://www.jaguar.com/
Official worldwide web site of Jaguar Cars. Directs users to pages tailored to specific markets and model-specific websites.

Jaguar HK
www.jaguar.com.hk/
Jaguar has always believed that a car is the closest thing you can create to a thing that is alive. Explore our range of luxury models, F-Type - XE - XF - XJ

Jaguar Cars - Wikipedia, the free encyclopedia
https://en.wikipedia.org/wiki/Jaguar_Cars
Jaguar Cars (/dʒəˈɡeɪər/ or JAG-ər) is a brand of luxury car manufactured by Jaguar, a multinational car manufacturer headquartered in Whitley, Coventry, England. Tata Motors - Jaguar XJ (X351) - Jaguar XF - Jaguar XJL - Jaguar XJ6 - Jaguar XJ8

Jaguar - Wikipedia, the free encyclopedia
The jaguar (/dʒəˈɡeɪər/) (Panthera onca) is a big cat, a feline in the Panthera genus, and is the only extant member of the genus Panthera native to the Americas. Jaguars are native to South America and Central America south to Argentina and Uruguay.
• Background
• Related work
• Our approach
• Experiments
• Summary
Heuristic approaches

- Maximal marginal relevance (MMR) criterion (Carbonell and Goldstein, SIGIR’98)
- Select documents with high divergence (Zhai et al., SIGIR’03)
- Minimize the risk of dissatisfaction of the average user (Agrawal et al., WSDM’09)
- Diversity by proportionality: an election-based approach (Dang and Croft, SIGIR’12)
- …
Heuristic approaches

• Maximal marginal relevance (MMR) criterion (Carbonell and Goldstein, SIGIR'98)
• Select documents with high divergence (Zhai et al., SIGIR'03)
• Minimize the risk of dissatisfaction of the average user (Agrawal et al., WSDM'09)
• Diversity by proportionality: an election-based approach (Dang and Croft, SIGIR'12)

Learning approaches

• SVM-DIV: formulate the task as a problem of predicting diverse subsets (Yue and Joachims, ICML’04)
• REC & RBA: online learning algorithms based on user’s clicking behavior (Radlinski et al., ICML’07)
• R-LTR: a process of sequential document selection and optimizing the likelihood of ground truth rankings (Zhu et al., SIGIR’14)
• …
Heuristic approaches

- Maximal marginal relevance (MMR) criterion (Carbonell and Goldstein, SIGIR'98)
- Select documents with high divergence (Zhai et al., SIGIR'03)
- Minimize the risk of dissatisfaction of the average user (Agrawal et al., WSDM'09)
- Diversity by proportionality: an election-based approach (Dang and Croft, SIGIR'12)

Learning approaches

- SVM-DIV: formulate the task as a problem of predicting diverse subsets (Yue and Joachims, ICML'04)
- REC & RBA: online learning algorithms based on user’s clicking behavior (Radlinski et al., ICML'07)
- R-LTR: a process of sequential document selection and optimizing the likelihood of ground truth rankings (Zhu et al., SIGIR'14)

Diversity evaluation measures

- Subtopic recall (Zhai et al., SIGIR’03)
- $\alpha$-NDCG (Clarke et al., SIGIR’08)
- ERR-IA (Chapella et al., CIKM’09)
- NRBP (Clarke et al., ICTIT’09)
- ...
Maximal marginal relevance (Carbonell and Goldstein, SIGIR'98)

$$MMR \overset{\text{def}}{=} \arg \max_{D_i \in R \setminus S} \left[ \lambda \left( \text{Sim}_1(D_i, Q) \right) - (1 - \lambda) \max_{D_i \in S} \text{Sim}_2(D_i, D_j) \right]$$

- **Advantage**
  - top-down user browsing behavior

- **Disadvantage**
  - non-learning: limited number of ranking signals
  - High parameter tuning cost
Formalization

Four key components: input space, output space, ranking function $f$, loss function $L$

$$\hat{f} = arg\min_{f \in F} \sum_{i=1}^{N} L(f(X^{(i)}, R^{(i)}), y^{(i)})$$
Relational Learning-to-Rank (Zhu et al., SIGIR’14)

- **Formalization**
  - Four key components: input space, output space, ranking function $f$, loss function $L$
    \[
    \hat{f} = \arg \min_{f \in F} \sum_{i=1}^{N} L(f(X^{(i)}, R^{(i)}), y^{(i)})
    \]

- **Definition of ranking function**
  \[
  f_S(x_i, R_i) = \omega_r^T x_i + \omega_d^T h_S(R_i), \forall x_i \in X \setminus S
  \]
  - relevance score
  - diversity score
Formalization

Four key components: input space, output space, ranking function $f$, loss function $L$

$$\hat{f} = \arg\min_{f \in F} \sum_{i=1}^{N} L(f(X^{(i)}, R^{(i)}), y^{(i)})$$

Definition of ranking function

$$f_S(x_i, R_i) = \omega_r^T x_i + \omega_d^T h_S(R_i), \forall x_i \in X \setminus S$$
• Definition of loss function

\[ L(f(X, R), y) = - \log P(y|X) \]

\[ P(y|X) = P(x_{y(1)}, x_{y(2)}, \ldots, x_{y(n)}|X) \]

• Plackett-Luce based Probability

\[ P(y|X) = \prod_{j=1}^{n} \frac{\exp \left\{ f_{S_{j-1}} (x_{y(j)}, R_{y(j)}) \right\}}{\sum_{k=j}^{n} \exp \left\{ f_{S_{k-1}} (x_{y(k)}, R_{y(k)}) \right\}} \]
• R-LTR Pros:
  • Modeling sequential user behavior in the MMR way
  • A learnable framework to combine complex features
  • State-of-the-art empirical performance

Can R-LTR be further improved?
Motivation

- R-LTR Cons:
  - Only utilizes “positive” rankings, but treat “negative” rankings equally
    - Not all negative rankings are equally negative (different scores)
    - How about using discriminative learning which is effective in many machine learning tasks?
  - Learning objective differs with diversity evaluation measures
    - How about directly optimizing evaluation measures?
learn MMR model using both positive and negative rankings

How to achieve this?

optimize diversity evaluation measures
• Background
• Related work
• Our approach
• Experiments
• Summary
Learning the ranking model

Basic loss function

\[
\hat{y}^{(n)} \text{ is the ranking constructed by the maximal marginal relevance model}
\]

\[
\min_{f_S} \sum_{n=1}^{N} L(\hat{y}^{(n)}, J^{(n)})
\]

\[
L(\hat{y}^{(n)}, J^{(n)}) \text{ is the function for judging the ‘loss’ of the predicted ranking } \hat{y}^{(n)} \text{ compared with the human labels } J^{(n)}
\]

\[
J^{(n)} \text{ denotes the human labels on the documents}
\]
Evaluation measures as loss function

- Aim to maximize the diverse ranking accuracy in terms of a diversity evaluation measure on the training data

\[
\sum_{n=1}^{N} \left( 1 - E(X^{(n)}, \hat{y}^{(n)}, J^{(n)}) \right)
\]

Difficult to directly optimize the loss as \( E \) is a non-convex function.

\( E \) represents the evaluation measures which measures the agreements between the ranking \( y \) over documents in \( X \) and the human judgements \( J \)
• Resort to optimize the upper bound of the loss function

\[
\sum_{n=1}^{N} \left( 1 - E(X^{(n)}, \hat{y}^{(n)}, J^{(n)}) \right)
\]

\(Y^+(n)\): positive rankings

\(Y^-(n)\): negative rankings

\(F(X, R, y)\): the query level ranking model

\[
F(X, R, y) = \text{Pr}(y|X, R)
= \text{Pr}(x_{y(1)} \cdots x_{y(M)}|X, R)
= \prod_{r=1}^{M-1} \text{Pr}(x_{y(r)}|X, S_{r-1}, R)
= \prod_{r=1}^{M-1} \frac{\exp\{f_{S_{r-1}}(x_{i,R_y(r)})\}}{\sum_{k=r}^{M} \exp\{f_{S_{r-1}}(x_{i,R_y(k)})\}}
\]

\(\llbracket \cdot \rrbracket\) is one if the condition is satisfied otherwise zero
Evaluation measures as loss function

- Resort to optimize the upper bound of the loss function

\[
\sum_{n=1}^{N} \left(1 - E(X^{(n)}, \hat{y}^{(n)}, J^{(n)})\right)
\]

\(Y^+(n)\): positive rankings

\[
\sum_{n=1}^{N} \max_{\begin{array}{c}
ym^+ \in Y^+(n) \\
ym^- \in Y^-(n)
\end{array}} \left( E(X^{(n)}, y^+, J^{(n)}) - E(X^{(n)}, y^-, J^{(n)}) \right) \cdot \left[ F(X^{(n)}, R^{(n)}, y^+) \leq F(X^{(n)}, R^{(n)}, y^-) \right]
\]

\(Y^-(n)\): negative rankings

Upper bounded if \(E \in [0, 1]\)

\(F(X, R, y)\): the query level ranking model

\[
F(X, R, y) = \Pr(y|X, R)
\]

\[
= \prod_{r=1}^{M-1} \Pr(x_{y(r)}|X, S_{r-1}, R)
\]

\[
= \prod_{r=1}^{M-1} \frac{\exp\{f_{S_{r-1}}(x_{i,Ry(r)})\}}{\sum_{k=r}^{M} \exp\{f_{S_{r-1}}(x_{i,Ry(k)})\}}
\]
Loss function can be optimized under the framework of Perceptron.

The algorithm is referred to as PAMM.

- Firstly, PAMM generates positive and negative rankings.
- Secondly, PAMM optimizes model parameters $\omega_r$ and $\omega_d$.

1. $\Delta F \leftarrow F(X^{(n)}, R^{(n)}, y^+) - F(X^{(n)}, R^{(n)}, y^-) $
2. if $\Delta F \leq E(X^{(n)}, y^+, J^{(n)}) - E(X^{(n)}, y^-, J^{(n)})$
3. then
4. calculate $\nabla \omega_r^{(n)}$ and $\nabla \omega_d^{(n)}$
5. $(\omega_r, \omega_d) \leftarrow (\omega_r, \omega_d) + \eta \times \left( \nabla \omega_r^{(n)}, \nabla \omega_d^{(n)} \right)$
6. end if

- Finally, PAMM outputs the optimized model parameters $(\omega_r, \omega_d)$. 

Direct optimization with Perceptron
Advantages of PAMM

• Adopting the ranking model that meets the maximal marginal relevance criterion

• Ability to direct optimize any diversity evaluation measure in training

• Ability to use both positive rankings and negative rankings in training
Outline

• Background
• Related work
• Our approach
• Experiments
• Summary
Experiment setting

- **Dataset**: TREC WT2009, WT2010, and WT2011

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#queries</th>
<th>#labeled docs</th>
<th>#subtopics per query</th>
</tr>
</thead>
<tbody>
<tr>
<td>WT2009</td>
<td>50</td>
<td>5149</td>
<td>3~8</td>
</tr>
<tr>
<td>WT2010</td>
<td>48</td>
<td>6554</td>
<td>3~7</td>
</tr>
<tr>
<td>WT2011</td>
<td>50</td>
<td>5000</td>
<td>2~6</td>
</tr>
</tbody>
</table>

- **Data processing**
  - Indri toolkit (version 5.2)
  - Porter stemmer and stopword removal

- **Evaluation**
  - TREC official measures: ERR-IA, $\alpha$-NDCG

- **Baselines**
  - QL, MMR, xQuAD, PM-2, ListMLE, SVM-DIV, StructSVM, R-LTR
Feature Vectors (Zhu et al., SIGIR’14)

- **Relevance features**
  - Weighing features: VSM, BM25, LM…
  - Term dependency features: MRF
  - Length
  - Pos
  - ...

- **Diversity features**
  - Cosine diversity
  - Jaccard diversity
  - Subtopic diversity
  - Document-level co-occurrence
  - …
Performance comparison of all methods

<table>
<thead>
<tr>
<th>Method</th>
<th>WT2009</th>
<th>WT2010</th>
<th>WT2011</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ERR-IA@20</td>
<td>α-NDCG@20</td>
<td>ERR-IA@20</td>
</tr>
<tr>
<td>QL</td>
<td>0.164</td>
<td>0.269</td>
<td>0.198</td>
</tr>
<tr>
<td>ListMLE</td>
<td>0.191</td>
<td>0.307</td>
<td>0.244</td>
</tr>
<tr>
<td>MMR</td>
<td>0.202</td>
<td>0.308</td>
<td>0.274</td>
</tr>
<tr>
<td>xQuAD</td>
<td>0.232</td>
<td>0.344</td>
<td>0.328</td>
</tr>
<tr>
<td>PM-2</td>
<td>0.229</td>
<td>0.337</td>
<td>0.330</td>
</tr>
<tr>
<td>SVM-DIV</td>
<td>0.241</td>
<td>0.353</td>
<td>0.333</td>
</tr>
<tr>
<td>StructSVM(ERR-IA)</td>
<td>0.261</td>
<td>0.373</td>
<td>0.355</td>
</tr>
<tr>
<td>StructSVM(α-NDCG)</td>
<td>0.260</td>
<td>0.377</td>
<td>0.352</td>
</tr>
<tr>
<td>R-LTR</td>
<td>0.271</td>
<td>0.396</td>
<td>0.365</td>
</tr>
<tr>
<td>PAMM(ERR-IA)</td>
<td>0.294</td>
<td>0.422</td>
<td>0.387</td>
</tr>
<tr>
<td>PAMM(α-NDCG)</td>
<td>0.284</td>
<td>0.427</td>
<td>0.380</td>
</tr>
</tbody>
</table>
Effects of positive and negative rankings

Ranking accuracies w.r.t. number of positive rankings

Ranking accuracies w.r.t. number of negative rankings
• New learning-to-rank model for search result diversification: PAMM
  ✓ Employs a ranking model that follows the maximal marginal relevance criterion
  ✓ Directly optimizes the diversity evaluation measures under the framework of Perceptron
  ✓ Ability to utilize both positive rankings and negative rankings in training

• Experimental results show that PAMM significantly outperforms the state-of-the-art baseline methods
THANKS!

Contact me:
Email: xialong@software.ict.ac.cn