

Learning to Rank を用いた QA サイトにおけるベストアンサーの推定

長野 敬介† 青野 雅樹†

† 豊橋技術科学大学 大学院 情報・知能工学専攻

1. はじめに

近年、ユーザ同士が質問や回答をやりとりして情報を得ることのできる Yahoo!知恵袋や教えて!goo などの質問応答サイト(QA サイト)における質問解答コンテンツ(以下 QA コンテンツ)には多くの質問が投稿され、各質問には回答が投稿される。その中から質問者はベストアンサーと呼ばれる、質問者にとっての最良の回答が決められる。しかし、質問に寄せられる回答は複数存在するものもあり質問者はその中から妥当だと考えられるベストアンサーを決めなくてはならない。そこで本研究では Learning to Rank を用いてベストアンサーを推定する研究を行う。ベストアンサーを推定する研究は過去にも行われており、西原ら[3]の研究では、質問と回答の文末表現を抽出し、機械学習によりベストアンサー推定が行われている。また、石川ら[4]は、複数判定者に手動で良質回答を判定してもらい、判定者らが良質回答の評価に用いた判定基準を特定するという研究を行っている。これらに対し本研究では「質問者はベストアンサーを選ぶ際、回答 1 つ 1 つを見比べる(ランク付けする)ことで一番良質な回答を選出している」という仮説を立て、Learning to Rank によるベストアンサー推定を行う。

2. Learning to Rank

Learning to Rank とは機械学習の一種である。Learning to Rank には、ランキングの対象となるデータに訓練データを用いて、数値スコアか順序スコアを定義する Pointwise 法、2 つのデータ間に相対的順序を定義する Pairwise 法、一連のデータ(リスト)に順序を定義する Listwise 法がある。本研究では前節で記述した仮説に最も合致した手法である Pairwise 手法を用いる。

3. 提案手法

本研究は統計的調査及び過去の研究の知見を基に大きく分けて学習に用いる素性、及びベストアンサー以外の回答に対するランキングモデルの 2 つを提案する。

3.1. 提案素性

ベストアンサー推定のために、回答及び回答者から得られる情報から、全素性 23 素性を学習する。

3.2. ランキングモデル

Learning to Rank では各素性に加え、適合度のランキングを行う。しかし、QA コンテンツではベストアンサー以

外の回答のランク付けがされていない。よって、本研究ではベストアンサー以外の回答に対して、回答及び回答者から得られる情報を基にランク付けを行う。

1. 時系列の観点から見たランキング(Time)

質問に対する回答の投稿時間が早い回答をランキング上位とするランキング手法。

2. 時系列の観点から見たランキングその 2(Time2)

質問に対する回答の投稿時間が遅い回答をランキング上位とするランキング手法。

3. テキスト情報量の観点から見たランキング(Length)

回答文のテキスト量及びリンクの有無を考慮したランキング手法

4. ユーザに印象付ける表現の観点から見たランキング(Emotion)

回答文内に存在するユーザに対して印象を与える形態素について考慮したランキング手法。

5. 回答文の事実性の観点から見たランキング(Fact)

回答文内に存在する曖昧な表現の有無を考慮したランキング手法。

6. 回答者の貢献度の観点から見たランキング(User)

回答者の過去の回答数やベストアンサー獲得率などを考慮したランキング手法。

以上 6 手法を用いて次章にて実験を行う。

4. 実験

4.1. 実験概要

Learning to Rank の有用性、及び提案した素性、及びランキングモデルの有用性を評価するため、評価実験を行った。実験データとして、国立情報学研究所[5]が提供している Yahoo!知恵袋データ(第 2 版)を用いた。

表 2. Yahoo!知恵袋データ(第 2 版)

期間	質問数	回答数
2004 年 4 月～2009 年 9 月	約 1600 万件	約 5000 万件

しかし 2004 年 4 月～2006 年 11 月までデータは Yahoo!知恵袋ベータ版からの移行の影響で時間情報に誤差が生じているため、本実験では対象外とした。対象カテゴリは「政治・社会問題」、「パソコン」、「恋愛相談、人間関係の悩み」、「Yahoo!オークション」とし、回答数が 5 件のもの限定して行った。表 3 に本実験で用いたデータ詳細を示す。本実験では質問 1000 件及びその回答 5000 件を 10 交差検定により評価する。推定に用いる判別器は RankingSVM を使い、提案素性及び回答のランキングを訓練する。実験内容としては 1. 各ランキングモデルの評価、2. 手作業によるランク付けを行った回答群を用いた評価、

Learning to Rank for Estimating the Best Answer from Q&A Web Sites

Keisuke Nagano Masaki Aono

† Dept. of Computer Science and Engineering, Toyohashi University of Technology

3. 素性選択を用いた後の推定精度の評価, 4. ランキングモデルの組み合わせを用いた後の推定精度の評価の全4実験を行う。評価尺度には P@1, MRR に加え, 実験 2 の手作業によるランク付けを行った回答群を用いた評価において nDCG を採用した。

表 3. 実験に使用したデータセット

(4 カテゴリー, 回答数 5 件の質問 1000 件を 10 交差検定)

データの種類	質問数	回答数
訓練データ	900 件	4500 件
テストデータ	100 件	500 件

4.2. 実験結果

実験 1 各ランキングモデルの評価

本実験では各ランキングモデルの評価を行う為, 表 3 のデータセットを用いて推定精度の評価を行った。ベースラインとしてランダムにランク付けしたモデルでの精度と, 提案素性を用いて通常の SVM を使った場合の精度, 石川らの研究[4]を用いた。なお用いた判別器は SVMlight となっている。実験の結果, P@1 及び MRR とともに時系列の観点から見たランキング手法その 2 とテキスト情報量の観点から見たランキング手法が各ベースラインよりも高評価であることを確認した。

表 4. 実験 1 結果(政治・社会問題カテゴリ)

Model	P@1	MRR
Time2	.540	.701
Length	.449	.607
Random	.205	.365
提案素性(SVM)	.281	.515
石川ら	.327	.556

実験 2 手作業によるランク付けを行った回答群を用いた評価

本実験ではベストアンサー以外の回答に対し手作業でランク付けを行った場合の推定精度の評価を行った。表 3 のデータセットの質問 100 件分に対する回答群を手作業によりランク付けを行った。評価内容は訓練データに手作業でランク付けしたものを加えた場合の P@1 及び MRR の評価とテストデータを手作業でランク付けしたデータとした場合の nDCG による評価である。実験の結果, 1 つ目の評価では実験 1 で高評価であった 2 ランキングモデルにおいて精度向上は見られなかった。また, 2 つ目の実験の nDCG による評価もランキングモデルごとに大きな差は見られなかった。これは質問者によってベストアンサー決定の決め手となる尺度が異なることが原因として考えられる。

実験 3 素性選択を用いた後の推定精度の評価

本実験では提案した全 23 素性の中で有用な素性のみを用いた場合の推定精度の評価を行った。予備実験として各素性を 1 つずつ抜いて推定した場合の推定結果の変化を調べた。これによりランキングモデルごとに有用な素性を確認し, これらのみを用いて推定を行った。これによ

り各ランキングモデルにおいて推定精度の向上を確認した。

実験 4 ランキングモデルの組み合わせを用いた後の推定精度の評価

本実験では提案したランキングモデルの組み合わせを用いた場合の推定精度の評価を行った。これまでの実験から得られた知見を基に各質問の回答群に対し事前に処理を行う。まずある質問のベストアンサーが Time2 モデルを当てはめた際に上位 n% である場合, その質問の回答群には Time2 モデルでランク付けを行う。そうでない場合, Length モデルを当てはめた際に上位 n% である場合, その質問の回答群には Length モデルを当てはめる。そうでない場合, Time2 モデルと Length モデルを組み合わせたモデル(Mix)によるランク付けを行う。さらに, 訓練に用いるデータに関して 1. 訓練データ全て 2. Time2 モデルを採用した回答群のみ 3. Length モデルを採用した回答群のみ 4. Mix モデルを採用した回答群のみの 4 パターンに関して推定精度の評価を行った。表 5 に図 1 の n=40, Time2 モデル及び Length モデルに関して素性選択を行った場合の推定精度を示す。

表 5. 実験 4 結果

(政治・社会問題カテゴリ, n=40, 素性選択後)

Model	P@1	MRR
Hybrid	.501	.671
Hybrid_only_Time2	.507	.678
Hybrid_only_Length	.362	.556
Hybrid_only_Mix	.681	.789

表 5 より Mix モデルを採用した回答群のみを訓練に用いた場合の精度が最も高評価であることを確認した。

5. おわりに

本稿では, 質問者のベストアンサー決定方法に仮説を立て, Learning to Rank を用いた Yahoo!知恵袋におけるベストアンサー推定を行った。また, 学習する際の素性及び, ベストアンサー以外の回答に対して仮ランキングを行うモデルを提案した。実験の結果 Learning to Rank によるベストアンサー推定の有用性を確認し, 質問者がどのような尺度でベストアンサーを決定しているかは質問者毎に異なるため, 質問毎に適切なランキングモデルを選択することで更なる精度向上につながることを確認した。

参考文献

- [1]Yahoo!知恵袋, <http://chiebukuro.yahoo.co.jp/>
- [2]教えて!goo, <http://oshiete.goo.ne.jp/>
- [3]西原陽子, 松村真宏, 谷内田正彦 “QA サイトにおける質問に適した回答の判定”, NLP 若手の会第 2 回シンポジウム, 2007
- [4]石川大介, 酒井哲也, 関洋平, 栗山和子, 神門典子, “QA サイトにおける質問に適した回答の判定”, 情報知識学会誌, Vol.20, No.2, 2010
- [5]国立情報学研究所(NII), <http://www.nii.ac.jp/>