Improving Question Retrieval in CQA with Label Ranking

Presented by WONG Wei

2010-10-5

Outline

- Problem & Motivation
- Label Ranking
- Improved QR Based on Label Ranking
- Experiment Design

Question Retrieval

Existing Methods of QR

- Vector Space Model

- Okapi BM25 Model

- Language Model

- Translation Model

- Translation-Based Language Model

An Example

What's the asker really wants to get from others?

Opinions?

Instructions?

Recommendation?

An Example

What's the asker really wants to get from others?

Recommendation?

How-to question?

Definition?

Motivation

- Existing methods does not consider users' different intent.

- Multi-label classification.

- Preferences among different labels.

Question Classification Based on User's Intention

Table 2: Types of questions

Type	Purpose	Expected answer
REC	Recommendation	Suggested item and the reason
FAC	Seeking facts	Data,location, or name
YNO	Yes/No decision	Yes/No and reason
HOW	How-to question	The instructions
WHY	Seeking reason	The explanation
TSL	Translation	Translated text in target language
DEF	Definition	The definition of given entity
OPN	Seeking opinion	Opinions of other users
TRA	Transportation	Navigation instructions
INT	Interactive	Discussion thread
MAT	Math problem	Solution with steps

- By Xiance Si et.al's work on VLDB 2010

Label Ranking (Formally)

- Given:
  - a set of training instances: $\{\mathbf{x}_k \mid k = 1 \ldots m\} \subseteq \mathbf{X}$
  - a set of labels $L = \{l_i \mid i = 1 \ldots n\}$
  - for each training instances $\mathbf{x}_k$: a set of pairwise preferences of the form $l_i \succ_{\mathbf{x}_k} l_j$ (for <u>some</u> of the labels)
- Find:

  a ranking function ($\mathcal{X} \to \Omega$ mapping) that maps each $\mathbf{x} \in \mathbf{X}$

  to a ranking $\succ_{\mathbf{x}}$ of L (Permutation $\pi_{\mathbf{x}}$) and generalizes well in terms of a loss function on rankings (e.g. , kendall′s tau coefficient)

Instanced Based Label Ranking

Nearest Neighbors
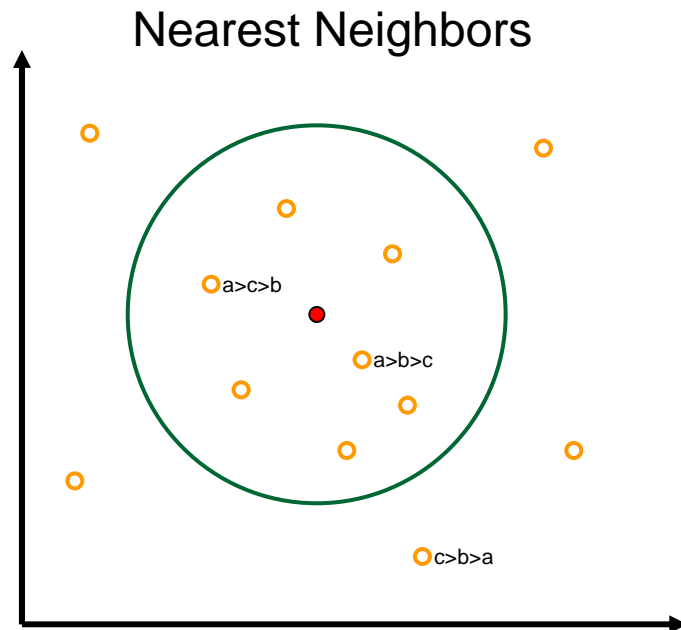
- Distribution of ranking is (approx.) constant in a local region.
- Core part is to estimate the locally constant model.

Instance Based Label Ranking

- Mallows model (Mallows, Biometrika, 1957)

$$\mathcal{P}(\sigma|\theta,\pi) = \frac{\exp(-\theta d(\pi,\sigma))}{\phi(\theta,\pi)}$$

- with

  center ranking $\pi \in \Omega$

  spread parameter $\theta > 0$

  and $d(\cdot)$ is a metric on permutations

  Computational issues arise when the training

  data contains incomplete rankings.

Instance Based Label Ranking

- Approximation via a variant of EM, viewing the non-observed labels as hidden variables.

1. Start with an initial center ranking (via generalized Borda count)
2. Replace an incomplete observation with its most probable extension (first M-step, can be done efficiently)
3. Obtain MLE as in the complete ranking case (second M-step)
4. Replace the initial center ranking with current estimation
5. Repeat until convergence

Probable Extension:

Observation σ	Extensions $E(\sigma)$
$a \_ b$	$a \_ b \_ c$ $a \_ c \_ b$ $c \_ a \_ b$

Improved QR Based on Label Ranking

- Linear combine label ranking with QR

$$S_{q,u} = \gamma CON_{q,u} + (1 - \gamma) INT_{q,u},$$

- With

  $CON_{q,u}$ is the content similarity,

  $INT_{q,u}$, is the intent similarity.

Experiment Design

- Data Analysis
  1. Number of relative question for each query
  2. Demonstrate the classification system degree
  3. Average number of labels for each question
- Classification methods for comparison:
  SVM & Naive Bayes & LR
- LR + Translation Model/Language Model
  /Translation-Based Language Model
- SVM+ Translation Model/Language Model
  /Translation-Based Language Model
- Naive Bayes+ Translation Model/Language Model
  /Translation-Based Language Model

Thanks!

Question and Comments?