A Generalized Framework of Exploring Category Information for Question Retrieval in Community Question Answer Archives

Xin Cao, Gao Cong, Bin Cui, and Christian S. Jensen

In *Proceedings of the 19th international conference on world wide web* (*WWW 2010*)

Prepared and Presented by Baichuan Li

Outline

- Introduction & Motivation
- Category-Enhanced Question Retrieval Models
- Experiments
- Conclusion

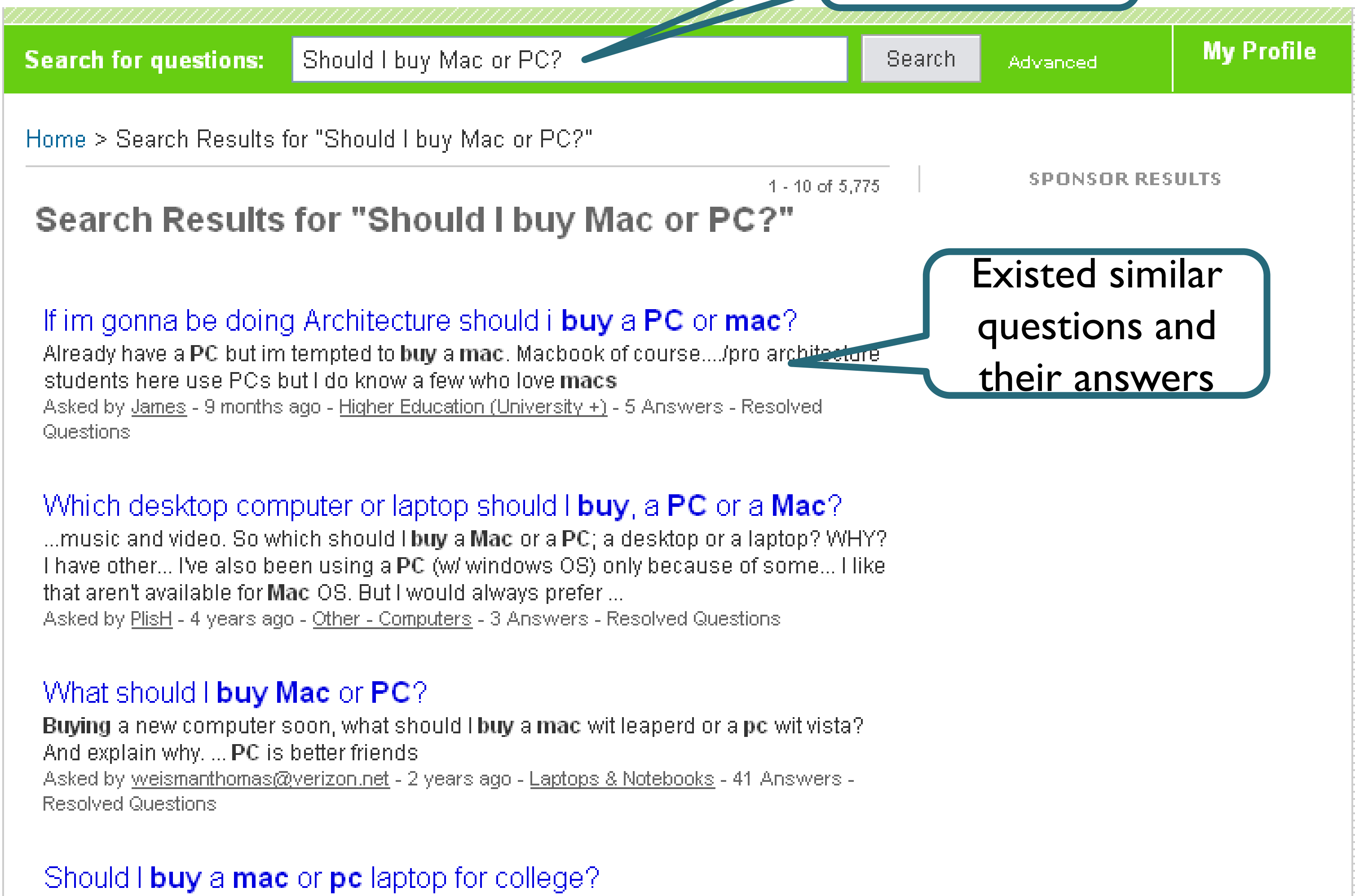
Introduction

- Community Question-Answering (CQA) Services

Baidu知道

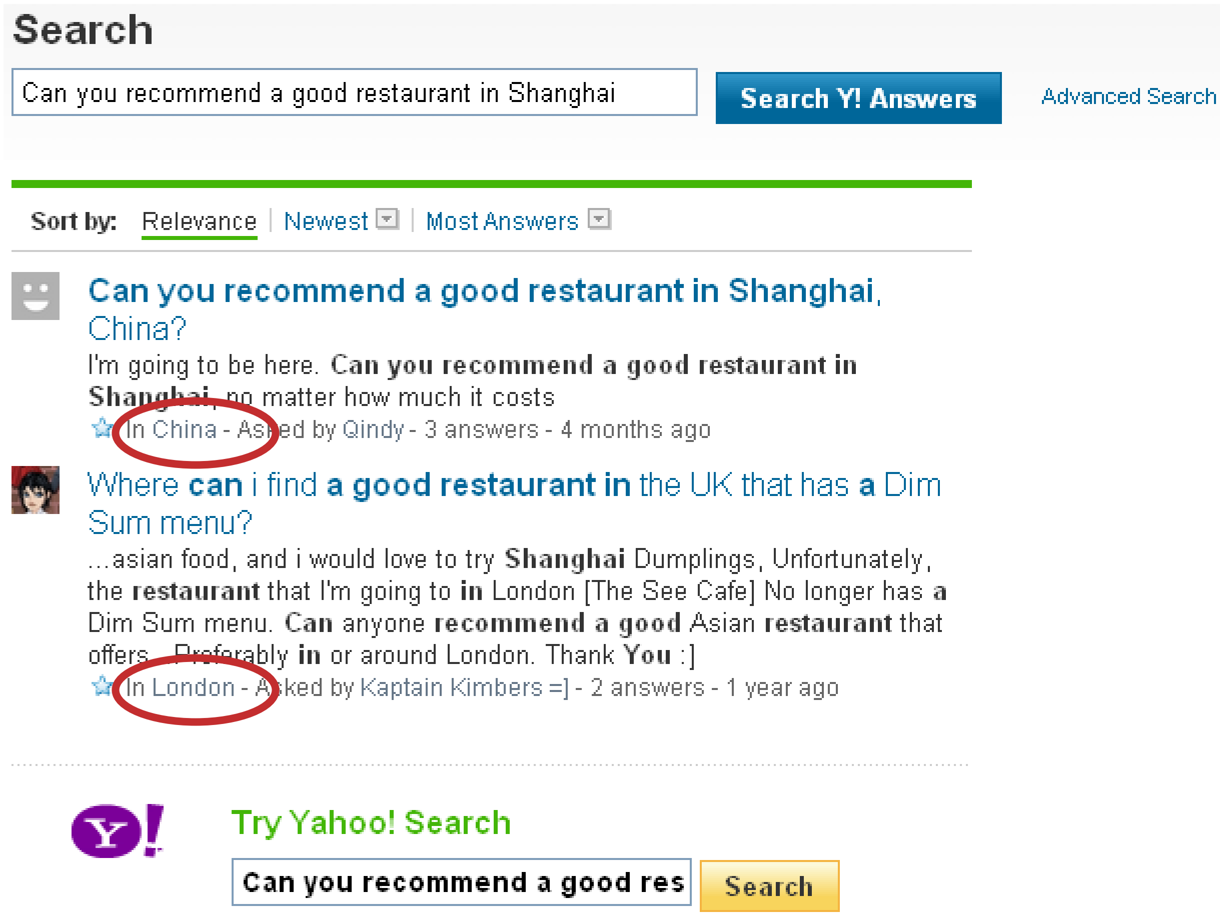
Question Retrieval

Query



Existed similar questions and their answers

Motivation



Query ⟷ Category

CATEGORY-ENHANCED QUESTION RETRIEVAL MODELS

Exploiting Categories in Question Retrieval

- Given a query **q**, a historical question **d**, and the category *cat*(**d**) that contains d:

$$RS_{\mathbf{q},\mathbf{d}} = (1-\alpha)N(S_{\mathbf{q},\mathbf{d}}) + \alpha N(S_{\mathbf{q},cat(\mathbf{d})})$$

where $S_{\mathbf{q},\mathbf{d}}$ is the local relevance score and $S_{\mathbf{q},cat(\mathbf{d})}$ is the global relevance score, N() is the normalization function and α is a weighting parameter.

- Words play different roles in computing local and global relevance scores

Retrieval Models

- Vector Space Model
- Okapi BM25 Model
- Language Model
- Translation Model
- Translation-Based Language Model

Vector Space Model

$$S_{\mathbf{q},\mathbf{d}} = \frac{\sum_{t \in \mathbf{q} \bigcap \mathbf{d}} w_{\mathbf{q},t} w_{\mathbf{d},t}}{W_{\mathbf{q}} W_{\mathbf{d}}}, \text{ where}$$

$$w_{\mathbf{q},t} = \ln(1 + \frac{N}{f_t}), \; w_{\mathbf{d},t} = 1 + \ln(tf_{t,\mathbf{d}})$$

$$W_{\mathbf{q}} = \sqrt{\sum_t w_{\mathbf{q},t}^2}, \; W_{\mathbf{d}} = \sqrt{\sum_t w_{\mathbf{d},t}^2}$$

Here N is the number of questions in the whole collection, f_t is the number of questions containing the term t, and $tf_{t,\mathbf{d}}$ is the frequency of term t in $\mathbf{d}$.

Vector Space Model

$$S_{\mathbf{q},\mathbf{d}} = \frac{\sum_{t\in\mathbf{q}\bigcap\mathbf{d}} w_{\mathbf{q},t} w_{\mathbf{d},t}}{W_{\mathbf{q}} W_{\mathbf{d}}}, \text{ where}$$

$$w_{\mathbf{q},t} = \ln(1 + \frac{N}{f_t}),\ w_{\mathbf{d},t} = 1 + \ln(tf_{t,\mathbf{d}})$$

$$W_{\mathbf{q}} = \sqrt{\sum_t w^2_{\mathbf{q},t}},\ W_{\mathbf{d}} = \sqrt{\sum_t w^2_{\mathbf{d},t}}$$

Global relevance score

$$S_{\mathbf{q},cat(\mathbf{d})} = \frac{\sum_{t\in\mathbf{q}\cap cat(\mathbf{d})} w_{\mathbf{q},t} w_{cat(\mathbf{d}),t}}{W_{\mathbf{q}}}, \text{ where}$$

$$w_{\mathbf{q},t} = \ln(1 + \frac{M}{fc_t}), w_{cat(\mathbf{d}),t} = 1 + \frac{1}{\ln(\frac{W_{cat(\mathbf{d})}}{tf_{t,cat(\mathbf{d})}})}$$

Here M is the total number of leaf categories, fc_t is the number of categories that contain the term t, $tf_{(t,cat(\mathbf{d}))}$ is the frequency of t in the category $cat(\mathbf{d})$, $W_{cat(\mathbf{d})}$ is the length of $cat(\mathbf{d})$ (number of words contained in $cat(\mathbf{d})$), and $w_{\mathbf{q},t}$ captures the IDF of word t with regard to categories.

Local relevance score

$$w_{\mathbf{q},t} = \ln(1 + \frac{N_{cat(\mathbf{d})}}{f_{t,cat(\mathbf{d})}})$$

Okapi BM25 Model

$$S_{\mathbf{q},\mathbf{d}} = \sum_{t \in \mathbf{q} \bigcap \mathbf{d}} w_{\mathbf{q},t} w_{\mathbf{d},t}, \text{ where}$$

$$w_{\mathbf{q},t} = \ln\left(\frac{N - f_t + 0.5}{f_t + 0.5}\right) \frac{(k_3 + 1) tf_{t,\mathbf{q}}}{k_3 + tf_{t,\mathbf{q}}}$$

$$w_{\mathbf{d},t} = \frac{(k_1 + 1) tf_{t,\mathbf{d}}}{K_{\mathbf{d}} + tf_{t,\mathbf{d}}}$$

$$K_{\mathbf{d}} = k_1\left((1 - b) + b \frac{W_{\mathbf{d}}}{W_A}\right)$$

Here N is the number of questions in the collection; f_t is the number of questions containing the term t; $tf_{t,\mathbf{d}}$ is the frequency of term t in $\mathbf{d}$; k_1, b, and k_3 are parameters.

Okapi BM25 Model

$$S_{\mathbf{q},\mathbf{d}} = \sum_{t \in \mathbf{q} \bigcap \mathbf{d}} w_{\mathbf{q},t} w_{\mathbf{d},t}, \text{ where}$$

$$w_{\mathbf{q},t} = \ln(\frac{N - f_t + 0.5}{f_t + 0.5}) \frac{(k_3 + 1) tf_{t,\mathbf{q}}}{k_3 + tf_{t,\mathbf{q}}}$$

$$w_{\mathbf{d},t} = \frac{(k_1 + 1) tf_{t,\mathbf{d}}}{K_{\mathbf{d}} + tf_{t,\mathbf{d}}}$$

$$K_{\mathbf{d}} = k_1((1 - b) + b \frac{W_{\mathbf{d}}}{W_A})$$

Global relevance score

$$S_{\mathbf{q},cat(\mathbf{d})} = \sum_{t \subset \mathbf{q} \bigcap cat(\mathbf{d})} w_{\mathbf{q},t} w_{cat(\mathbf{d}),t}, \text{ where}$$

$$w_{\mathbf{q},t} = \ln(\frac{M - fc_t + 0.5}{fc_t + 0.5}) \frac{(k_3 + 1) tf_{t,\mathbf{q}}}{k_3 + tf_{t,\mathbf{q}}}$$

$$w_{cat(\mathbf{d}),t} = \frac{(k_1 + 1) tf_{t,cat(\mathbf{d})}}{K_{\mathbf{d}} + tf_{t,cat(\mathbf{d})}}$$

$$K_{\mathbf{d}} = k_1((1 - b) + b \frac{W_{cat(\mathbf{d})}}{W_{A(cat)}})$$

Local relevance score

$$w_{\mathbf{q},t} = \ln(\frac{N_{cat(\mathbf{d})} - f_{t,cat(\mathbf{d})} + 0.5}{f_{t,cat(\mathbf{d})} + 0.5}) \frac{(k_3 + 1) tf_{t,\mathbf{q}}}{k_3 + tf_{t,\mathbf{q}}}$$

$$K_{\mathbf{d}} = k_1((1 - b) + b \frac{W_{\mathbf{d}}}{W_{A,cat(\mathbf{d})}})$$

Language Model

$$S_{\mathbf{q},\mathbf{d}} = \prod_{t \in \mathbf{q}} ((1-\lambda)P_{ml}(t|\mathbf{d}) + \lambda P_{ml}(t|\mathbf{Coll})), \text{ where}$$

$$P_{ml}(t|\mathbf{d}) = \frac{tf_{t,\mathbf{d}}}{\sum_{t' \in \mathbf{d}} tf_{t',\mathbf{d}}}$$

$$P_{ml}(t|\mathbf{Coll}) = \frac{tf_{t,\mathbf{Coll}}}{\sum_{t' \in \mathbf{Coll}} tf_{t',\mathbf{Coll}}}$$

Here $P_{ml}(t|\mathbf{d})$ is the maximum likelihood estimate of word t in $\mathbf{d}$; $P_{ml}(t|\mathbf{Coll})$ is the maximum likelihood estimate of word t in the collection $\mathbf{Coll}$; and λ is the smoothing parameter.

Language Model

$$S_{\mathbf{q},\mathbf{d}} = \prod_{t \in \mathbf{q}} ((1-\lambda) P_{ml}(t|\mathbf{d}) + \lambda P_{ml}(t|\mathbf{Coll})), \text{ where}$$

$$P_{ml}(t|\mathbf{d}) = \frac{tf_{t,\mathbf{d}}}{\sum_{t' \in \mathbf{d}} tf_{t',\mathbf{d}}}$$

$$P_{ml}(t|\mathbf{Coll}) = \frac{tf_{t,\mathbf{Coll}}}{\sum_{t' \in \mathbf{Coll}} tf_{t',\mathbf{Coll}}}$$

Global relevance score

d -> Cat(**d**)

Local relevance score

Coll -> Cat(**d**)

Translation Model

$$S_{\mathbf{q},\mathbf{d}} = \prod_{t \in \mathbf{q}} \left((1-\lambda) \boxed{\sum_{w \in \mathbf{d}} T(t|w)} P_{ml}(w|\mathbf{d}) + \lambda P_{ml}(t|\mathbf{Coll}) \right)$$

$T(t|w)$ denotes the probability that word w is the translation of word t.

IBM translation models:

http://en.wikipedia.org/wiki/Statistical_machine_translation

Translation Model

$$S_{\mathbf{q},\mathbf{d}} = \prod_{t \in \mathbf{q}} ((1-\lambda) \sum_{w \in \mathbf{d}} T(t|w) P_{ml}(w|\mathbf{d}) + \lambda P_{ml}(t|\mathbf{Coll}))$$

Global relevance score

d -> Cat(**d**)

Local relevance score

Coll -> Cat(**d**)

Translation-Based Language Model

$$S_{\mathbf{q},\mathbf{d}} = \prod_{t\in\mathbf{q}}((1-\lambda)(\beta\sum_{w\in\mathbf{d}}T(t|w)P_{ml}(w|\mathbf{d}) + (1-\beta)P_{ml}(t|\mathbf{d})) + \lambda P_{ml}(t|\mathbf{Coll}))$$

β controls the translation component's impact.

Global relevance score

d -> Cat(**d**)

Local relevance score

Coll -> Cat(**d**)

EXPERIMENTS

Data Set

- Question Repository

Category	Question#	Category	Question#
Arts & Humanities	114737	Health	183181
Beauty & Style	49532	Home & Garden	50773
Business & Finance	154714	Local Businesses	69581
Cars & Transportation	208363	News & Events	27884
Computers & Internet	129472	Pets	72265
Consumer Electronics	126253	Politics & Government	85392
Dining Out	58980	Pregnancy & Parenting	63228
Education & Reference	107337	Science & Mathematics	116047
Entertainment & Music	196100	Social Science	61011
Environment	28476	Society & Culture	122358
Family & Relationships	53687	Sports	275893
Food & Drink	55955	Travel	403926
Games & Recreation	72634	Yahoo! Products	228368

- Query Set
  - 252 queries from http://homepages.inf.ed.ac.uk/gcong/qa

Results

	VSM	OptC	QC	VSM+VSM	%chg	Okapi+VSM	%chg	LM+VSM	%chg	TR+VSM	%chg	TRLM+VSM	%chg
MAP	0.2407	0.2414	0.2779	**0.3711**	54.2*	0.3299	37.1*	0.3632	50.9*	0.3629	50.8*	0.3628	50.7*
MRR	0.4453	0.4534	0.4752	**0.5637**	26.6*	0.5314	19.3*	0.5596	25.7*	0.5569	25.1*	0.5585	25.4*
R-Prec	0.2311	0.2298	0.2568	**0.3419**	48.0*	0.3094	33.9*	0.3366	45.7*	0.3346	44.8*	0.3357	45.3*
P@5	0.2222	0.2289	0.2436	**0.2789**	25.5*	0.2559	15.2*	0.2746	23.6*	0.2746	23.6*	0.2753	23.9*

Table 1: VSM vs. CE with VSM for computing local relevance (%chg denotes the performance improvement in percent of each model in CE; * indicates a statistically significant improvement over the baseline using the t-test, p-value < 0.05)

	Okapi	OptC	QC	VSM+Okapi	%chg	Okapi+Okapi	%chg	LM+Okapi	%chg	TR+Okapi	%chg	TRLM+Okapi	%chg
MAP	0.3401	0.2862	0.3622	0.4007	17.8*	0.3977	16.9*	**0.4138**	21.7*	0.4082	20.0*	0.4132	21.5*
MRR	0.5406	0.4887	0.5713	0.6131	13.4*	0.5884	8.8	0.6214	15.0*	0.6172	14.2*	**0.6215**	15.0*
R-Prec	0.3178	0.2625	0.3345	0.3648	14.8*	0.3613	13.7*	0.3758	18.3*	0.3677	15.7*	**0.3762**	18.4*
P@5	0.2857	0.2824	0.2998	0.3140	9.9*	**0.3176**	11.2*	0.3161	10.6*	0.3111	8.8	0.3147	10.2*

Table 2: Okapi vs. CE with Okapi for computing local relevance (%chg denotes the performance improvement in percent of each model in CE; * indicates a statistically significant improvement over the baseline using the t-test, p-value < 0.05)

Conclusion

- Exploiting category information associated with questions for improving question retrieval
- Conducting experiments with large scale CQA data
- Improvements
  - Considering answers
  - Utilizing hierarchical category structures
  - …