Intent-Aware Query Similarity

Jiafeng Guo[1], Xueqi Cheng[1], Gu Xu[2], Xiaofei Zhu[1]
[1]Institute of Computing Technology, CAS, China
[2]Microsoft Bing

Outline

- Motivation
- Our Approach
- Experimental Results
- Conclusions

Motivation

- Query Similarity Calculation: Key element of various IR applications
  - query recommendation
  - query expansion
  - advertisement matching
  - ...
- Properly define the similarity is Non-Trivial
  - High ambiguity: multiple potential search intent

Motivation

Apple

Apple tree

search intent:
looking for apple fruits

Apple store

search intent:
find products of the apple company

Intent-aware query similarity

Similarity between queries defined upon search intent

Existing Methods

Intent-Not-Aware Intent-Aware

Pare-wise Measures

Independent measured on each pair

Jaccard coefficient [Beeferman et al. 2000]
cosine similarity [Baeza-Yates et al. 2004; Wen et al. 2002]
Hybrid methods [Zhang et al. 2006; Jones et al. 2006]
Jaccard & cosine [Deng et al. 2009]
Kernel method [Sahami et al. 2006]

Mixed representation
Biased by popular intent
Ignore unpopular ones

Apple ~ Apple store
Apple ≠ Apple tree

Graph-based Measures

Propagate similarity over query relation graph

Random walk [Craswell et al. 2007]
hitting time [Mei et al. 2008]
SimRank [Antonellis et al. 2008]
Matrix Factorization [Ma et al. 2008]
Graph Projection [Bordino et al. 2010]

Propagate across the boundary
Wrongly connect queries from different search intents

Apple store ~ Apple tree

Outline

- Motivation
- Our Approach
- Experimental Results
- Conclusions

Main Idea

A. Identify the potential search intent of queries

B. Intent-aware similarity measure

I. Extract intent-aware representations

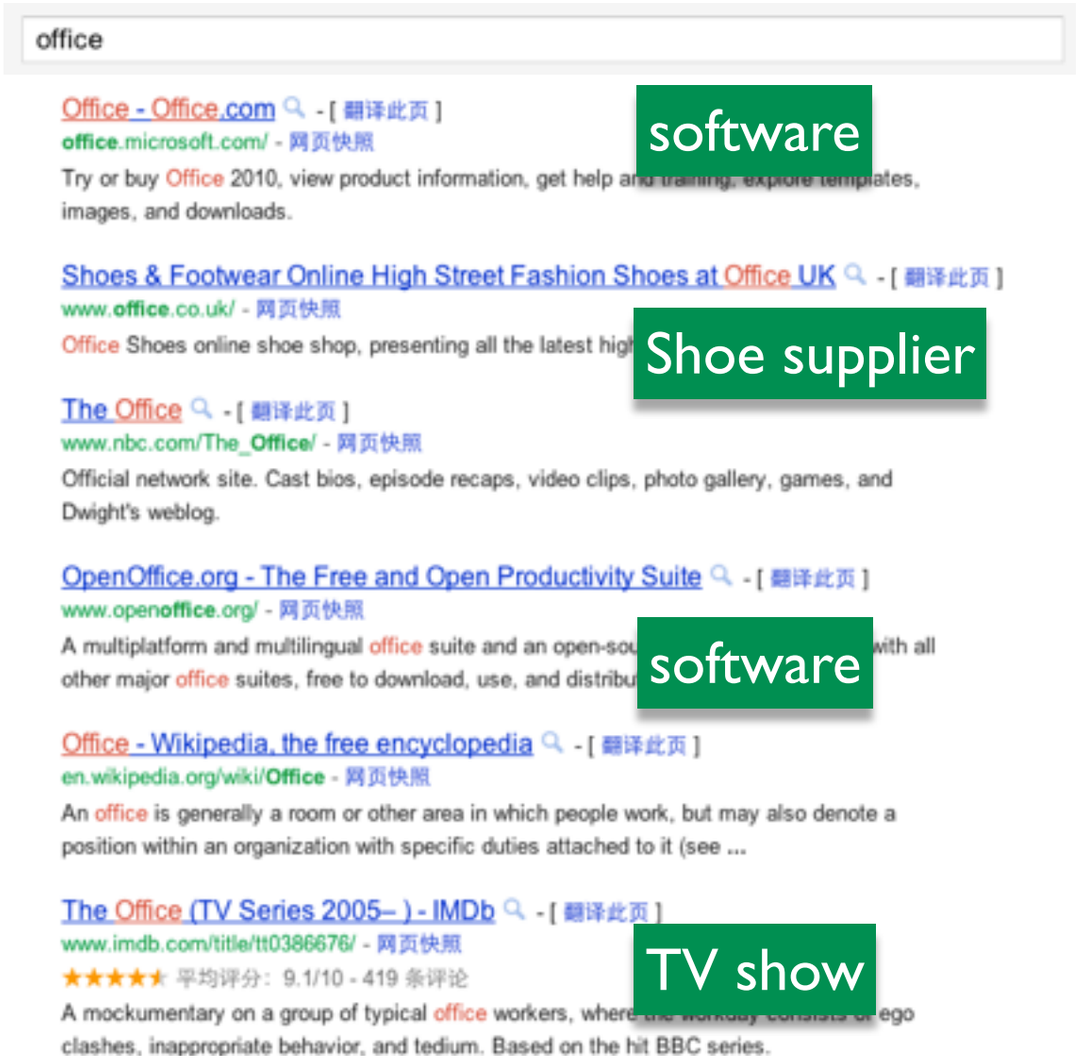
II. Apply different types of similarity measures

A. Identify Search Intents (Data)

leverage two types of auxiliary data

Search result snippets

Great Context Describing the query

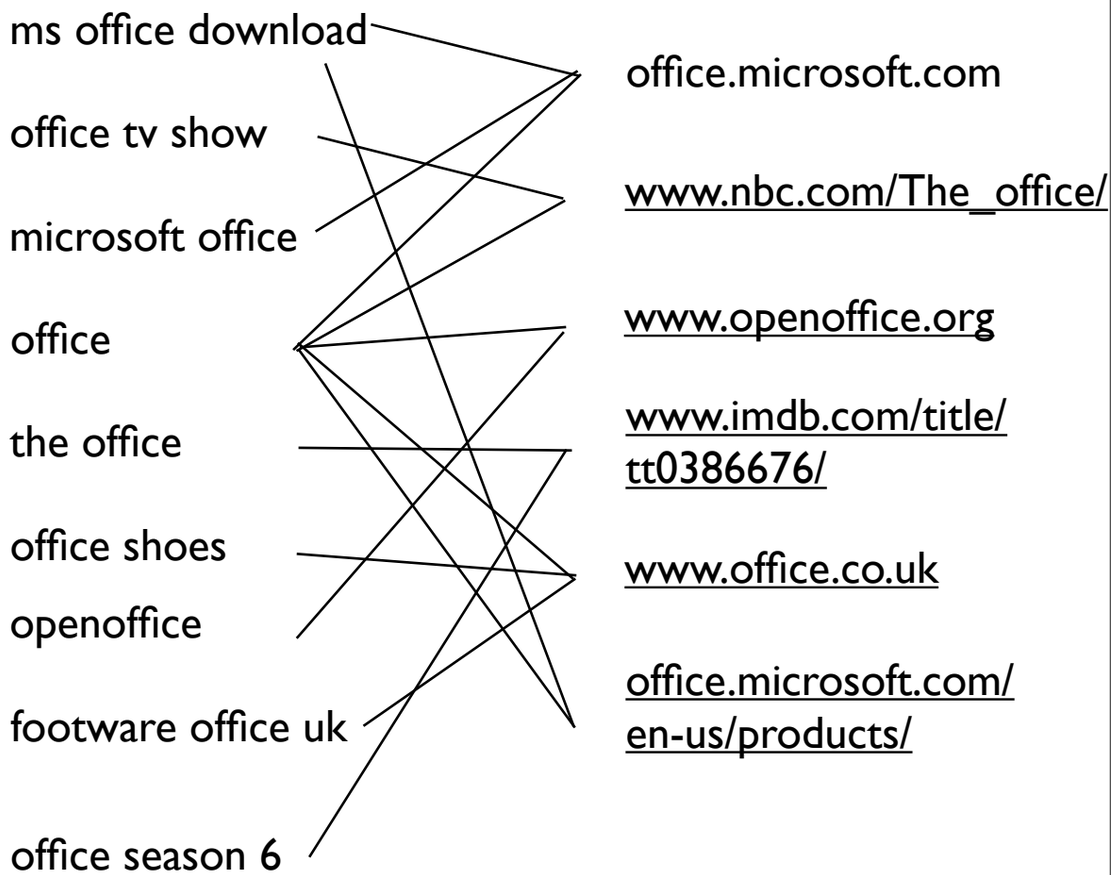


Pro: higher recall

Con: irrelevant/spam/advertisement/ambiguity

Clickthrough

Precise information from Wisdom of crowds



Pro: higher precision

Con: sparse

A. Identify Search Intents (Algorithm)

Search result snippets

Topic Model

top search result snippets → virtual documents

words in snippets → words

potential search intents → topics

PLSI model

1. select a query q_i with probability $P(q_i)$,
2. pick a potential search intent s_k with probability $P(s_k|q_i)$
3. generate a word w_j with probability $P(w_j|s_k)$.

log-likelihood

$$\tilde{\mathcal{L}} = \sum_{i=1}^{N}\sum_{j=1}^{M} n(q_i, w_j) \log\left(P(q_i) \sum_{k=1}^{K} P(w_j|s_k) P(s_k|q_i) \right)$$

Clickthrough

Regularization

powerful constraint: two queries share many same clicked URLs → convey similar search intent

$$\mathcal{R} = \sum_{i,j=1}^{N}\sum_{k=1}^{K} C_{ij} (P(s_k|q_i) - P(s_k|q_j))^2$$

co-click matrix

Regularized Topic Model

$$\begin{aligned} \mathfrak{L} &= \mathcal{L} - \lambda \mathcal{R} \\ &= \sum_{i=1}^{N}\sum_{j=1}^{M} n(q_i, w_j) \log\left(P(q_i) \sum_{k=1}^{K} P(w_j|s_k) P(s_k|q_i) \right) - \lambda \sum_{i,j=1}^{N}\sum_{k=1}^{K} C_{ij} (P(s_k|q_i) - P(s_k|q_j))^2 \end{aligned}$$

A. Identify Search Intents (Learning)

Generalized EM algorithm

E-step:

posterior probabilities $P(s_k|q_i,w_j)=\frac{P(w_j|s_k)P(s_k|q_i)}{\sum_{k'=1}^{K}P(w_j|s_{k'})P(s_{k'}|q_i)}$

M-step:

maximize the expected complete data log-likelihood

$$Q(\Phi,\Theta)=Q_1(\Phi,\Theta)-\lambda Q_2(\Theta)$$
$$=\sum_{i=1}^{N}\sum_{j=1}^{M}n(q_i,w_j)\sum_{k=1}^{K}P(s_k|q_i,w_j)\log[P(w_j|s_k)P(s_k|q_i)]\ -\lambda\sum_{i,j=1}^{N}\sum_{k=1}^{K}C_{ij}(P(s_k|q_i)-P(s_k|q_j))^2$$

$$P(w_j|z_k)=\frac{\sum_{i=1}^{N}n(q_i,w_j)P(s_k|q_i,w_j)}{\sum_{j'=1}^{M}\sum_{i=1}^{N}n(q_i,w_{j'})P(s_k|q_i,w_{j'})}$$

$$P(s_k|q_i)_{n+1}^{(t+1)}=(1-\gamma)P(s_k|q_i)_{n+1}^{(t)}+\gamma\frac{\sum_{j=1}^{N}C_{ij}P(s_k|q_j)_{n+1}^{(t)}}{\sum_{j=1}^{N}C_{ij}}$$

B.Intent-Aware Similarity Measure (Pair-wise)

Similarity independently measured by pair-wise metrics

I. Extract intent-aware representations

original: word vector representation $\vec{q}_i[l] = n(q_i, w_l)$

intent-aware: word vector representation under k-th search intent $\vec{q}_{ik}[l] = n(q_i, w_l)P(s_k|q_i, w_l)$

expected search intent distribution for each word occurrence w_l given query q_i

II. Apply Pair-wise similarity measures

similarity under k-th search intent $Sim_k(q_i, q_j) = \frac{\vec{q}_{ik} \cdot \vec{q}_{jk}}{\| \vec{q}_{ik} \| \| \vec{q}_{jk} \|}$

B.Intent-Aware Similarity Measure (Graph-based)

similarity calculated over the query graph

I. Extract intent-aware representations

original: query similarity graph adjacency matrix

$$A = [W_{ij}]_{i,j=1,\ldots,N}$$

Jaccard coefficient

intent-aware: the probability that an edge will be generated between query q_i with search intent s_k and query q_j with search intent s_l

$$P(s_k|q_i)P(s_l|q_j)$$

$$\sum_{k,k'} P(s_k|q_i)P(s_l|q_j) = 1$$

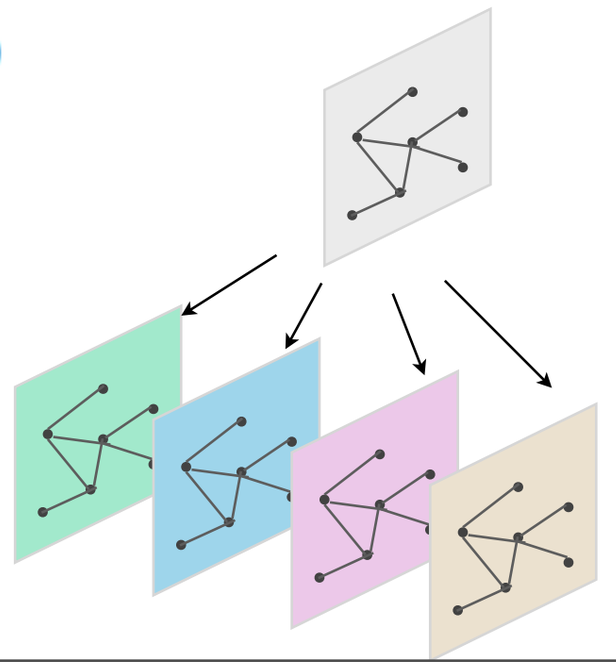
query similarity graph under k-th search intent $W_{ij}^k = W_{ij}P(s_k|q_i)P(s_k|q_j)$

II. Apply Graph-based similarity measures

spectral embedding $L_k\mathbf{y} = \lambda D_k\mathbf{y}$

query representation under k-th search intent $\vec{q}_{ik} = (\mathbf{y}_1(i), \ldots, \mathbf{y}_m(i))$

similarity under k-th search intent $Sim_k(q_i, q_j) = \frac{1 + cos(\vec{q}_{ik}, \vec{q}_{jk})}{2}$



Outline

- Motivation
- Our Approach
- Experimental Results
- Conclusions

Experiment Setting

- Data set:
  - one month sampled query logs from a commercial search engine
  - top 10 search results from the same search engine
  - 11,524 unique queries; 87,415 unique URLs; 45,882 unique words
- Baselines:
  - Intent-not-aware measures:
    - pair-wise (**Cos-Word**): cosine similarity based on tf-Idf weighted word vector
    - graph-based (**Embed-Click**): spectral embedding over the similarity graph based on clickthrough
  - Intent-aware measures:
    - pair-wise: **Cos-Intent**
    - graph-based: **Embed-Intent**

Qualitative Evaluation

Example Queries Pairs with Similarity Scores Calculated by Different Methods

Method	Intent†	taylor					
		taylor swift	taylor swift new songs	taylor ice cream	taylor soft serve machine	taylor acoustic	taylor guitars
Cos-Word	N/A	0.55	0.51	0.49	0.58	0.62	0.59
Cos-Intent	singer	0.76	0.68	0	0	0	0
	instrument	0	0	0	0	0.87	0.85
	company	0	0	0.52	0.61	0	0
Embed-Click	N/A	0.48	0.47	0.47	0.46	0.44	0.51
Embed-Intent	singer	1	1	0	0	0	0
	instrument	0	0	0	0	0.60	0.63
	company	0	0	0.87	0.72	0	0

†the search intents are manually labeled for illustration

Qualitative Evaluation

Examples of Similar and Dissimilar Query Pairs

Type	Query Pair	Traditional Method		Intent-Aware Method†	
		Cos-Word	**Embed-Click**	**Cos-Intent**	**Embed-Intent**
Similar Pairs	(apple, apple store)	0.86	0.89	0|0.92	0|1
	(apple, apple fruit)	0.17	0.46	0.44|0	0.83|0
Dissimilar Pairs	(apple store, apple fruit)	0.09	0.37	0|0	0|0
	(apple ipod, apple tree)	0.08	0.34	0|0	0|0
Similar Pairs	(taylor, taylor swift)	0.55	0.48	0.76|0|0	1|0|0
	(taylor, taylor soft serve machine)	0.58	0.46	0|0|0.61	0|0|0.72
Dissimilar Pairs	(taylor swift, taylor soft serve machine)	0.28	0.36	0|0|0	0|0|0
	(taylor ice cream, taylor acoustic)	0.24	0.38	0|0|0	0|0|0

†similarity scores under different intents are separated by vertical bars for clarity

Quantitative Evaluation

Ground truth for evaluation:
manually label similar queries under each major intent for a set of test queries
totally 1,581 labeled queries

Examples of Manually Built Test Set

Seed Query	Major Intents
24	1. tv show 24, 24 on fox, 24 the series 2. 24 fitness, 24hr fitness, 24 hour gym
sigma	1. sigma aldrich, sigma chemicals, sigma biology 2. greek alphabet sigma, sigma symbol, sigma maths 3. sigma camera, sigma photo, sigma lenses
svm	1. svm cards, svm gift card, svm gas cards 2. svm kernel, svm tutorial, support vector machine

Expected Inter-intent Similarity:

$$InterSim(S)=\frac{1}{K(K-1)}\sum_{S_k,S_{k'}\in S,k\neq k'}\left[\sum_{q_i\in S_k}\sum_{q_j\in S_{k'}}\frac{Sim(q_i,q_j)}{|S_k||S_{k'}|}\right]$$

Expected Intra-intent Similarity:

$$IntraSim(S)=\frac{1}{K}\sum_{k=1}^{K}\left[\sum_{q_i,q_j\in S_k,i\neq j}\frac{2Sim(q_i,q_j)}{|S_k||S_k-1|}\right]$$

Expected inter-intra ratio $\mathcal{H}_{\hat{S}}(Sim)=E\left[\frac{InterSim(S)}{IntraSim(S)}\right]_{S\in\hat{S}}$

$\mathcal{H}_{\hat{S}}(Sim)$ for Different Similarity Measures

Method	$\mathcal{H}_{\hat{S}}(Sim)$	Significant differences†
Cos-Word	0.47±0.06	>Embed-Click***
Cos-Intent	0.08±0.03	>Cos-Word*** >Embed-Click***
Embed-Click	0.54±0.02	
Embed-Intent	0.09±0.03	>Cos-Word*** >Embed-Click***

†the significant levels are denoted as 0.1* 0.05 ** 0.01 ***

Evaluation of Topic Models

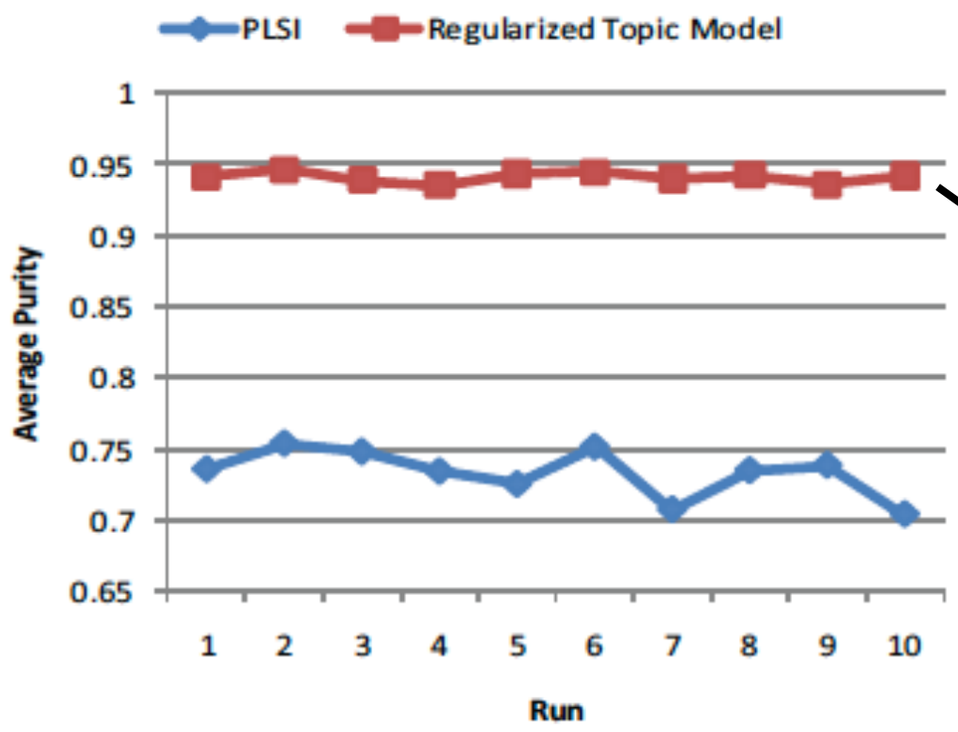
Does the regularization from query co-clicks really helps for the learning problem?

intent groups predicted $S = \{s_1, \ldots, s_J\}$

intent groups labeled $\hat{S} = \{\hat{s}_1, \ldots, \hat{s}_K\}$

Topic model learns better if the predicted intent groups are more like the human labeled results

$$Purity(S, \hat{S}) = \frac{1}{N} \sum_j \max_K |s_j \cap \hat{s}_k|$$



Higher purity score means better prediction on intent groups

Application

Query Recommendation

Structured Query Recommendation: diverse recommendation to enhance users' click behavior

query: iphone

iphone 3g	[iphone related]
apple iphone	iphone 3g
iphone price	iphone price
iphone review	iphone review
unlock iphone	unlock iphone
iphone plans	iphone apps
iphone jailbreak	[apple product]
iphone apps	ipod touch
iphone ringtones	mobileme
iphone verizon	[smartphones]
	blackberry
	palm
	nexus one

It is natural to apply intent-aware similarity measures for structured query recommendation

Evaluation on Query Recommendation

List approach: Cos-Word Structured approach: Cos-Intent

Evaluation Metric

Clicked Recommendation Number (CRN) $CRN_q = |\{r_i | l_i > 0, i \in [1,k]\}|$

Clicked Recommendation Score (CRS) $CRS_q = \frac{\sum_{i=1}^{k} l_i}{CRN_q}$

Total Recommendation Score (TRS) $TRS_q = \frac{\sum_{i=1}^{k} l_i}{k}$

} click willingness

Comparisons between List Approach and Our Approach on Click Performance

	List Approach	Our Approach
Ave. CRN	4.10	4.63 (+12.9%)
Ave. CRS	0.43	0.47 (+9.3%)
Ave. TRS	0.15	0.17 (+13.3%)

Utility and effectiveness of our intent-aware approach in real applications

Outline

- Motivation
- Our Approach
- Experimental Results
- Conclusions

Conclusions

- As the first attempt, we cast some light on the problem of "intent-aware query similarity"
- Measure similarity with respect to search intent
  - A regularized topic model to identify search intents using snippets and co-clicks
  - Extract query representation under different intents
  - Apply different types of similarity with intent-aware representation
- Experiments demonstrate the effectiveness of our measure
- Future work
  - Using more context information for identify search intents
  - Apply intent-aware query similarity in other real applications

Acknowledgement

- National High-tech R&D Program of China under grant No. 2010AA012500
- National Natural Science Foundation of China under grant No. 61003166 and No. 60933005

THANKS!