【VLDB2015 & SIGIR2015 勉強会】

Terms-1: Learning to Reweight Terms with Distributional Representations

担当：欅惇志（東工大）

Learning to Reweight Terms with Distributional Representations,Guoqing Zheng, Jamie Callan (CMU)

- 概要
  - 文書中の語の重みではなく，(クエリ) 語に重みを付与

直感的には TF の役割　　直感的には IDF の役割

  - (クエリ) 語の重みは term recall weight を使用
    - 適合文書中に出現する語の重みが大きくなるように語の重み付与
  - term recall weight が最大になるように語の重み学習する
    - 語の feature は分散表現 (word vector) から射影

※ 図表の一部は，論文及び東北大岡崎先生の WebDB Forum 2015 のチュートリアル資料から引用
http://www.slideshare.net/naoakiokazaki/ss-55494101

Distributional representation (分散表現)

- word2vec の有名な例
  - vec(King) – vec(Man) + vec(Woman) = vec(Queen)

- 分布仮説 (distributional memory/semantics)
  - 似た意味/属性の語同士は似たコンテキスト中で出現

… packed with people drinking **beer** or wine. Many restaurants …
into alcoholic drinks such as **beer** or hard liquor and derive …
… in miles per hour, pints of **beer**, and inches for clothes. M…
…ns and for pints for draught **beer**, cider, and milk sales. The
carbonated beverages such as **beer** and soft drinks in non-ref…
…g of a few young people to a **beer** blast or fancy formal part…
…c and alcoholic drinks, like **beer** and mead, contributed to a…
People are depicted drinking **beer**, listening to music, flirt…
… and for the pint of draught **beer** sold in pubs (see Metricat…

… ith people drinking beer or **wine**. Many restaurants can be f…
…gan to drink regularly, host **wine** parties and consume prepar…
principal grapes for the red **wines** are the grenache, mourved…
… four or more glasses of red **wine** per week had a 50 percent …
…e would drink two bottles of **wine** in an evening. According t…
…. Teran is the principal red **wine** grape in these regions. In…
…a beneficial compound in red **wine** that other types of alcohol
… Colorino and even the white **wine** grapes like Trebbiano and …
In Shakesperean theatre, red **wine** was used in a glass contai…

beer と wine の周辺の語彙はそこそこ似てる？

Distributional representation (分散表現)

- word2vec の有名な例
  - vec(King) – vec(Man) + vec(Woman) = vec(Queen)
- 分布仮説 (distributional memory/semantics)
  - 似た意味/属性の語同士は似たコンテキスト中で出現
- 単語の分散表現 (word vector 作成)
  - beer={1, 1, 0, …}
  - wine={1,1,0,…}
  - car={0,0,0,…}
  - beer と wine の内積は大きく, beer と car の内積が小さくなるように学習 (CBOW, Skip-gram)

複合語は各 word vector の合成で表現可能
vec(Chinese river) = vec(Chinese) + vec(river)

クエリ vec も, 各クエリ語 vec の合成でよい

提案手法の概要

- (クエリ) 語の重みは term recall weight を使用
  - 適合文書中に出現する語の重みが大きくなるように語の重み付与
    - 教師あり学習 (適合文書ラベル)
    - BM25 やクエリ尤度モデルを拡張
    - ランキング学習のようなもの?
  - どの語が出現すると適合文書である確率が高い?
    - unlike IDF
  - 既存研究では人手による feature selection 必要

- term recall weight が最大になるように語の重みを学習
  - 語の feature は分散表現 (word vector) から射影
  - feature selection (パラメタチューニング?) は自動

Query Model	Language Model Retrieval			
	GOV2		ClueWeb09B	
	NDCG@20	ERR@20	NDCG@20	ERR@20
BOW	0.3732	0.1493	0.1597	0.1253
SD	0.4059	0.1607	0.1694	0.1243
DeepTR-SD (GOV2)	0.4121^{b}	0.1626^{b}	0.1743^{b}	0.1312_{s}
DeepTR-SD (ClueWeb09B)	0.4146^{b}	0.1614^{b}	0.1663	0.1255
DeepTR-SD (Google)	0.4112^{b}	0.1600	0.1698	0.1302

Query Model	BM25 Retrieval			
	GOV2		ClueWeb09B	
	NDCG@20	ERR@20	NDCG@20	ERR@20
BOW	0.3789	0.1487	0.1188	0.1075
SD	0.3940	0.1554	0.1294	0.1059
DeepTR-SD (GOV2)	0.4008_{s}^{b}	0.1590_{s}^{b}	0.1307	0.1068
DeepTR-SD (ClueWeb09B)	0.4033_{s}^{b}	0.1587_{s}^{b}	0.1305	0.1067
DeepTR-SD (Google)	0.4010_{s}^{b}	0.1586_{s}	0.1298	0.1050

b : Statistically significant difference with BOW
s : Statistically significant difference with SD

提案手法は非 term recall モデルと比較して精度向上