

## 融合广告主行为的拍卖词实时触发

解忠乾<sup>1\*</sup>, 常笑<sup>2</sup>, 姬东鸿<sup>1</sup>

(1. 武汉大学 计算机学院, 武汉 430072; 2. 百度在线网络技术(北京)有限公司, 北京 100085)

(\* 通信作者电子邮箱 doumoxiao@163.com)

**摘要:**搜索引擎触发广告的过程中,需要实时计算拍卖词(Bidword)和用户查询(Query)的相关性,广告语境下的Term动态赋权方式和短语商业价值评估成为相关性计算必须考虑的问题。为此引入广告主行为,结合连续词袋模型(CBOW),提出了一种广告语境下的短语相关计算方法ADPCB。首先通过CBOW模型获得短语中每个Term的向量;然后分析广告主行为,构建关于短语的全局赋权树,对短语结构进行分析得到Term的动态权重;最后将Term权重和向量线性组合产生短语的向量表示,用于Bidword和Query的相关性度量。对10 000对带有标签的Query和Bidword(正负比例1:1)利用Word2vec进行实验,ADPCB比结合CBOW模型的TF-IDF效果更好;而在准确率达到0.70时,ADPCB比潜在狄利克雷分布(LDA)、BM25和TF-IDF获得了更高的召回率。结果表明ADPCB提高了触发Bidword和Query的相关性,同时可以量化短语中Term的商业价值属性,减少低商业价值Query的广告触发数量,可应用于实时计算的场景。

**关键词:**广告触发;相关性;行为分析;词向量;商业价值

**中图分类号:** TP391 **文献标志码:** A

### Real-time advertising trigger with advertiser behavioral analysis

XIE Zhongqian<sup>1\*</sup>, CHANG Xiao<sup>2</sup>, JI Donghong<sup>1</sup>

(1. School of Computer, Wuhan University, Wuhan Hubei 430072, China;

2. Baidu Online Network Technology (Beijing) Company Limited, Beijing 100085, China)

**Abstract:** In the process of advertising on search engines, it needs to calculate the correlation between auction word (Bidword) and user's query (Query) in real time. Dynamic Term weight in advertisements and phrase commercial value assessment must be considered in relevant calculation. Thus, a phrase related calculation approach named ADPCB was proposed based on behavioral analysis and Continuous Bag-Of-Words (CBOW) model to deal with those problems. Firstly, this approach got vector of each Term by CBOW. Secondly, to analyze advertiser's behavior and construct a global empowerment tree about phrases, the phrase structure was analyzed to obtain dynamic Term weight. Finally the phrase distributed representation produced by Term weight and linear combination was applied to the related measurement between Bidword and Query. Experiments were conducted on 10000 pairs Query and Bidword (positive and negative ratio is 1:1) with editorial judgments by using Word2vec. ADPCB performed better than Term Frequency-Inverse Document Frequency (TF-IDF) which combined with CBOW; when the accuracy was 0.70, ADPCB got higher recall than that of Latent Dirichlet Allocation (LDA), BM25 (Best Match25) and TF-IDF. The experimental results and analysis show that ADPCB can recognize the commercial value quality of the phrase to reduce the quantity of advertising trigger of low commercial value Query, it can be used in real-time calculation scene.

**Key words:** advertising trigger; correlation; behavioral analysis; distributed representation of word; commercial value

## 0 引言

在线广告是互联网企业的主要收入来源,目前其市场规模高达数百亿美元,因此计算广告技术蓬勃发展。所谓计算广告是根据给定的用户及上下文,通过计算得到与之最匹配的广告,并进行精准定向投放的广告机制。采用该机制可以大幅度地提高广告主投放广告的点击率(Click Through Rate, CTR),增加广告所投放网站的访问量,帮助用户获取优质信息<sup>[1]</sup>。搜索引擎广告是计算广告中的重要方式,最终广告的展现主要依据用户查询(Query)。搜索引擎对用户的搜索请

求提供自然结果的同时,会在指定位置显示商业广告,并通过点击进行计费。搜索引擎为了获取最大的收益同时保证用户的搜索满意度,需要控制广告和Query的相关性。搜索引擎系统中通过触发对广告进行相关性过滤,本文主要研究拍卖词触发模块,包括评估Query的商业价值,对Query和拍卖词(Bidword)的相关性进行评分,过滤拍卖词。

广告中的Query平均长度为2.2个词,包含的信息很少,是相关性计算研究中的一个难点。目前主流的文本相关性计算主要基于向量空间模型<sup>[2]</sup>、主题模型、词典和句子分析方法,例如文献[3]中基于《知网》和二部图最大权匹配算法

收稿日期:2014-04-02;修回日期:2014-05-21。 基金项目:国家自然科学基金资助项目(61133012,61070082)。

作者简介:解忠乾(1989-),男,山东菏泽人,硕士研究生,CCF会员,主要研究方向:自然语言处理、数据挖掘;常笑(1982-),男,吉林四平人,博士研究生,主要研究方向:计算广告学、数据挖掘;姬东鸿(1967-),男,湖北武汉人,教授,CCF会员,主要研究方向:自然语言处理。

进行词汇的语义相似度计算,充分利用了语义信息,但是对于短语效果有限;文献[4]使用语义依存进行句子相似度计算,是一种深层结构分析法,时间复杂度较高,不适合广告中的实时触发。上述的相关性计算方法在广告语境效果不理想,因为 Query 和 Bidword 都是短语,信息量不足容易导致部分匹配转义问题,使得 Query 相关 Bidword 被过滤,文献[5]中使用自动 Query 扩展(Query expansion),将短语扩展为短文本。在短语扩展基础上,利用潜在狄利克雷分布(Latent Dirichlet Allocation, LDA)等主题模型可以计算短语相关性<sup>[6]</sup>,在语义层面保持相关性的同时,通过主题匹配度,弥补了基于关键词方法的不足,更好地利用了上下文信息,但细粒度需求上区分能力较差,可能会展现需求漂移大的广告<sup>[7]</sup>。文献[8-9]基于神经网络和层次化概率语言模型提出了 CBOW(Continuous Bag-Of-Words)模型,能够把词语转化为  $K$  维向量,利用向量空间的相关性表示其语义相关性,可以灵活地用于同义词判断、词汇聚类 and 词性分析。CBOW 模型结合 TF-IDF(Term Frequency-Inverse Document Frequency)赋权可以对短语进行相关性判断,但 TF-IDF 仅仅反映 Term 在资源中的出现概率<sup>[10]</sup>,是一种静态权值,无法获得相同 Term 在不同语境下重要性的变化,而且对于 Term 的商业价值无法进行评估,使得某些商业性很低的 Query 触发大量广告,影响了用户的搜索体验。

本文借鉴之前的成果,结合搜索引擎广告的具体情况提出了一种短语相关计算方法 ADPCB(Advertisement Phrase Correlation with Behavior Analysis),它既可以评估短语的商业价值,又能够动态计算 Term 权值,基于 CBOW 模型,计算拍卖词触发中 Query 和 Bidword 的相关性,进行拍卖词过滤,提高广告系统中用户查询和拍卖词的相关性。

## 1 相关理论

### 1.1 搜索引擎广告机制

为了控制相关性,目前大部分的搜索引擎都是通过下面三个步骤来展现广告:1)根据用户请求检索到一定数量的广告;2)对这批广告进行 CTR 预估并根据广告的出价和 CTR 进行排序;3)决定要展现广告的数量和位置,在搜索结果中进行展现<sup>[11]</sup>。

在过程1)中,搜索引擎广告主要采用了两类广告检索方法:确切匹配和高级匹配。采用确切匹配对广告进行检索时,通常是将广告数据以记录的形式保存在数据库中,并将拍卖词作为广告记录属性的一部分,依赖于数据库的查询处理机制来实现检索。采用高级匹配方法来对广告进行检索,通常是将广告数据以文档的形式保存在文档库中,然后根据用户查询利用信息检索技术来对广告进行检索<sup>[1]</sup>。

为了方便广告检索,大多数的搜索引擎通过拍卖词触发广告。广告主在购买词后需要将自己的广告和已购拍卖词关联,触发时首先根据用户查询选择拍卖词,然后使用拍卖词触发对应的广告,需要严格控制拍卖词和广告的相关性。拍卖词触发建立在广告库的基础上,精准的触发和广告库质量相关。

### 1.2 拍卖词触发

拍卖词触发是投放机制中对拍卖词进行筛选的方法,主要基于 Query 分析,包括 Query 改写、专名识别、Query 组成成

分分析、Query 扩展等。拍卖词触发流程如图1所示,右侧为每一步的举例。

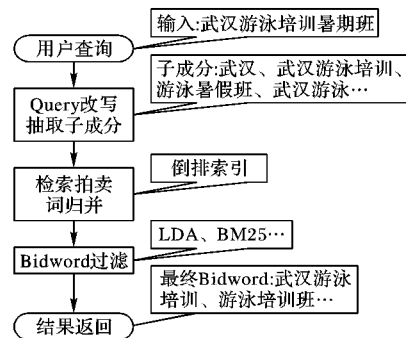


图1 拍卖词触发流程

在 Query 改写过程中,需要对用户点击行为进行挖掘,分析 Query 查询意图。此过程是基于 Term 对齐的同义变换技术:每个 Term 有若干个同义改写片段,所有同义改写片段构成一个图,同义变换问题转化成在图上寻找最优路径的问题。

Query 改写后,通过枚举方式抽取子成分,如 Query:ABC,子成分集合为{ABC,AB,BC,AC,A,B,C}。之后进行有效性检验:假如子成分 IDF 大于 IDF 最低阈值,同时在拍卖词索引中,则为有效子成分,由此建立起 Query 和子成分对应的 Bidword 之间的触发关系,得到一系列的候选 Bidword。本文将主要研究如何在候选 Bidword 中进行相关性过滤<sup>[12]</sup>,得到最终的 Bidword 列表。

## 2 短语扩展和 CBOW 模型

### 2.1 短语扩展

在拍卖词触发时,针对短语信息量不足导致的部分匹配转义问题,采用伪相关反馈方法对短语进行离线扩展。首先使用短语查询搜索引擎获得自然结果,选取前面的若干条摘要,然后统计 TF-IDF 较高的 Term,并使用相似度工具进行校验,从共现程度、独立表意能力、和原词的相关性方面进行过滤得到质量高的 Term,将这些 Term 作为短语的扩展。例如,“便宜飞机票”扩展出的部分结果如下:飞机票特价机票、机票、便宜、打折机票、国际机票、去哪儿、更便宜机票查询、国航、机票预订、航班查询……

LDA 主题模型可以运用于文本语义相似度计算,针对短语需要首先进行 Query 扩展。在后面的实验部分,先对短语进行扩展,然后使用 LDA、BM25(Best Match25)作为基准方法。

BM25 算法是二元独立模型的扩展,通常用于搜索相关性评分,在本文方法中使用它对 Query 和 Bidword 进行相关计算。BM25 方法首先对 Query 进行语素解析,然后计算搜索结果和每个语素的相关性评分,最后将相关性得分加权得到一个搜索结果和 Query 的相关性。BM25 的计算公式如下:

$$S(Q, D) = \sum_{i=1}^n w(q_i) \cdot \frac{f_i(k_1 + 1)}{f_i + k_1 \left(1 - b + \frac{b \cdot dl}{avgdl}\right)} \frac{qf_i(k_2 + 1)}{qf_i + k_2} \quad (1)$$

其中:  $f_i$  为  $q_i$  在  $D$  中的出现频率,  $qf_i$  为  $q_i$  在 Query 中的出现频率,  $k_1$ 、 $k_2$ 、 $b$  为调节因子,  $dl$  为文档  $D$  的长度,  $avgdl$  为所有文

档的平均长度。

## 2.2 CBOW 模型

CBOW 模型是建立在神经网络和语言模型基础上一种将词表征为  $K$  维向量的高效模型。单机版 CBOW 模型每天训练词汇数量为千亿级别,采用异步随机梯度下降训练大规模深度网络的方法,可以实现其分布式版本,对大规模数据进行快速的线下训练。

CBOW 使用的神经网络框架在 Hierarchical NNLM 的基础上去掉了隐藏层<sup>[13]</sup>,而且根据词频建立哈夫曼树,有效降低了计算量。图2是CBOW计算时采用的框架,可看出当前词的计算基于其上下文语境<sup>[14]</sup>。

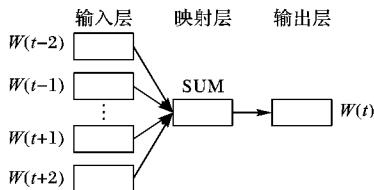


图2 CBOW 框架

CBOW 采用 Hierarchical softmax 对模型进行优化。其中  $v = (v_1, v_2, \dots, v_L) (v_j \in (0, 1))$  是当前词  $V$  的哈夫曼编码,  $V$  的上下文  $(w_{t-n+1}, \dots, w_{t-1}, w_{t+1}, \dots, w_{t+n-1})$ , 简记为  $context$ , 当前最终要优化的目标是:

$$p(v | context) = \prod_{j=1}^L p(v_j | v_1, v_2, \dots, v_{j-1}, context)^{1-v_j} \cdot (1 - p(v_j | v_1, v_2, \dots, v_{j-1}, context))^{v_j}$$

其中:

$$p(v_j | v_1, v_2, \dots, v_{j-1}, context) = \text{sigmoid}(\mathbf{c}(context) \cdot \mathbf{c}(v_j))$$

$$\mathbf{c}(v_j) \cdot \mathbf{c}(context) = \sum_{j \in context} \mathbf{c}(w_j)$$

$\mathbf{c}(v_j)$  为  $v_j$  所在哈夫曼树节点上的词向量。使用 CBOW 得到的是关于词指定长度的向量,可以使用这组向量衡量词语之间的相关性。

## 3 广告中的短语相关计算

本文提出的 ADPCB 本质上是一种组合语义方法,通过建立词语向量间的关系模型,把它们进行组合来合成短语的向量表示,可以灵活地利用各种信息。ADPCB 方法主要基于 CBOW 模型和 Term 动态赋权。2.2 节中已介绍了 CBOW 模型,以下将介绍 Term 动态赋权和 ADPCB 计算短语相关值的流程。

### 3.1 基于广告主行为的 Term 赋权

在搜索引擎广告系统中,广告主会购买大量的拍卖词,这些词和广告主产品相关,所以同一广告主购买的词语具有一定的关联;同时,拍卖词可以重复地被不同广告主购买。利用广告主的这些行为对短语进行结构分析获得 Term 的权重,可以一定程度上表征 Term 的商业价值。这个方法充分利用广告主的购买行为,关注短语部分和整体的相关性度量。本文使用全局赋权树模型得到部分与整体的关系。

使用全局赋权树进行训练首先有一个基本假设:较长的短语可以使用更短的短语进行近似。在广告库中可以用购买次数和共同购买次数来体现:Bidword 被购买次数越多越重

要;Bidword 的子串与 Bidword 共同购买得越多,两者的相关性越强。这样一个 Bidword 与其子串可以构建一个树,通过分析节点和边的关系,对其子成分进行赋权。

**定义1** 节点。在全局赋权树中一个节点对应一个短语或者 Term。

**定义2** 子短语。由短语内若干 Term 按照顺序组合而成的短语。

**定义3** 全局赋权树。以短语本身为根节点,子短语为非根节点,包含关系为边构成的树,当子短语为叶子节点时,对应为单 Term。

**定义4** 层次下移。在全局赋权树中,从根节点开始,计算相邻两层的子节点在父节点权重后,下移一层。例如首先计算第2层节点在第1层中父节点权重,然后计算第3层在第2层父节点中权重。

**定义5** 节点静态权值。静态权值计算公式如下:

$$f(v) = w_{f_v} / (w_{f_v} + a) \quad (2)$$

其中:  $w_{f_v}$  是拍卖词被购买次数;  $a$  是平滑系数,根据广告库规模调整。

**定义6** 边。节点存在包含关系则存在边,  $e_{ij}$  为拍卖词间的共同购买次数,  $b$  是系数,边的权值计算公式如下:

$$g(v_i, v_j) = e_{ij} / (e_{ij} + b) \quad (3)$$

在上述定义基础上,自底向上地推导父节点中某个 Term 的权值。

$$w(t) = \sum_{i=1}^n w_{v_i}(t) f(v_i) g(v_i, v_r) \quad (4)$$

假如父节点为  $v_r$ , 在树的  $m$  层,那么  $w_{v_i}(t)$  是叶子节点在  $m+1$  层父节点  $v_i$  的权重,  $f(v_i)$  和  $g(v_i, v_r)$  是由式(2)、(3)计算的结果,这两项乘积为子节点在上一层父节点的权重。

在 Term 动态赋权前,首先需要遍历广告库,使用哈希表存储构建全局赋权树所需的信息,表中每个节点包括短语的购买次数以及子短语共同被购买的次数。通过查询哈希表可以更快地建立全局赋权树获得 Term 动态权重,计算过程如下所示:

1) 将要进行 Term 赋权的短语作为根节点,构建全局赋权树的同时计算相邻两层子节点的权重,由式(2)~(4)计算得到子节点的权重。

2) 层次下移,计算相邻两层的子节点权重,方法和1)相同。通过上述方法计算得到所有相邻层的子节点在父节点的权重,至此赋权树构建完毕,每个节点中存储它在上一层父节点的权重。

3) 遍历叶子节点,使用式(4)自底向上顺序计算叶子节点在上层节点中的权重,直到上层节点为根节点,最后短语内部 Term 权值归一化,得到每个叶子节点在根节点的权重。

例如:在图3中,Term“雅思”在短语“武汉雅思培训学校”中的权值由“武汉雅思”“雅思培训”“雅思学校”中的“雅思”权值共同决定。计算得到第2层3个节点在根节点“武汉雅思培训学校”中的权值分别为  $\alpha$ 、 $\beta$ 、 $\gamma$ , 同理得到“雅思”在这三个节点的权重为  $m$ 、 $n$ 、 $q$ 。那么,“雅思”在“武汉雅思培训学校”中的权重就是:

$$w(\text{“雅思”}) = a \cdot m + \beta \cdot n + \gamma \cdot q$$



同理计算得到其他叶节点在根节点中的权重,最后归一化。

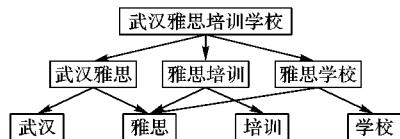


图3 Term 赋权示例

### 3.2 Query 商业价值量化

采用广告主行为进行 Term 赋权,短语中的 Term 权值既可以代表 Term 在短语中的重要性,也可以表达其自身的商业价值属性。短语在归一化前对所有 Term 权重加和可一定程度上反映短语的商业价值,使用它可以控制一些低商业价值 Query 广告触发数量。例如,用户搜索“周润发电影有哪些”,它属于资讯类 Query,广告主购买此类 Query 数量少,导致节点静态权值和边的权值较小,最终得到的未归一化 Term 权重之和会很低。利用此值可以减少此类 Query 的广告触发数量,甚至不显示广告,提高用户的搜索满意度。

### 3.3 ADPCB 方法流程

ADPCB 方法主要基于 CBOW 模型和 Term 动态赋权。在 2.2 节和 3.1 节基础上,计算 Query 和 Bidword 的相关值  $Sim(q, b)$  过程如下:

1) 对 Query 构建全局赋权树,利用 3.1 节中算法得到 Term 的权重  $W_1 = [w_{11}, w_{12}, \dots, w_{1m}]$ 。

2) 使用 TF-IDF 获得 Term 的权重  $W_2 = [w_{21}, w_{22}, \dots, w_{2m}]$ ,利用下式获得 Term 的最终权重:

$$W = \alpha W_1 + (1 - \alpha) W_2 \quad (5)$$

3) 使用 CBOW 模型获得各 Term 的向量表示  $[V_1(t), V_2(t), \dots, V_m(t)]$ , Query 的向量表示如下:

$$V(S) = \sum_{i=1}^m W_i(t) V_i(t)$$

4) 重复 1) ~ 3), 获得 Bidword 的向量表示。

5) 利用余弦距离得到  $Sim(q, b)$ 。

## 4 实验和分析

### 4.1 实验数据和评估标准

短语相关性是一个比较抽象的概念,目前在搜索引擎广告中对方法的评估主要通过人工标注语料计算准确率和召回率。实验使用的测试数据为 10 000 对带有标签的 Query 和 Bidword,正负比例 1:1。获取测试数据过程如下:首先通过线下广告日志挖掘准备标注的 Query 和 Bidword 短语对,进行数据清洗和过滤;然后由专业的标注人员遵循统一的标准对挖掘的数据进行标注;最后从标注语料中随机选择 10 000 对测试数据。

为了保证标注数据的可靠性,标注人员采用统一的标准,弱相关短语对标注为负例值为 0,例如:“西安幼儿英语培训”“英语三级培训”值为 0。负例主要包括 5 类:

1) 超集词:搜索词所涵盖的业务范围大于搜索结果。示例 Query 和 Bidword:PE 管件、PE 波纹管。

2) 语义、意图不相符:搜索词与搜索结果在语义、需求和业务等方面存在不同角度的不相关。示例 Query 和 Bidword:苹果 4 越狱、苹果 4。

3) 地域不一致:搜索词与搜索结果均含有地域词,但地域不一致。示例 Query 和 Bidword:重庆吃喝玩乐、吃在成都。

4) 语义宽泛词:搜索词语义过于宽泛,涵盖内容过多意图不明确。示例 Query 和 Bidword:青浦区、找工作。

5) 问题词:搜索词中存在明显存在咨询信息意图,商业价值小。示例 Query 和 Bidword:信用卡最低还款额是什么意思、信用贷款。

强相关短语对标注为正例值为 1,例如:“当下最赚钱的生意”“最新致富项目”值为 1。正例主要包括 3 类:

1) 子集词:搜索词业务是搜索结果业务的一部分。示例 Query 和 Bidword:养鸡、养殖。

2) 语义相关词:搜索词与搜索结果意思相同或相近。示例 Query 和 Bidword:手机价格、手机多少钱。

3) 意图相关词:Query 与 Bidword 表示相同需求或行业。示例 Query 和 Bidword:5 至 6 元童装、5 元童装加盟。

广告检索中相关性计算最重要的指标是召回率和准确率,在实验中,准确率  $P$ 、召回率  $R$  和  $F$  度量值分别定义如下:

$$P = \text{标记正例中正确数量} / \text{标记为正例数量}$$

$$R = \text{标记正例中正确数量} / \text{语料中正例数量}$$

$$F1 = 2PR / (P + R)$$

### 4.2 实验结果评估和分析

Word2vec 是 Google 于 2013 年发布的基于 CBOW 模型的开源词向量工具,使用效果和训练数据的质量、数量有很大的关系,选取广告库中 2 周的展现日志,抽取出用户点击的广告 Query 和 Bidword,进行数据清洗和过滤,得到的训练数据大约为 2 亿条。单独使用 TF-IDF 结合 CBOW 模型对测试数据进行评估,之后加入 3.1 节中方法,即采用 ADPCB,两者进行对比。

在实验中采用的向量维度  $K = 200$ ,在 Term 动态赋权中,式(2)、(3)中参数  $a, b$  为 2,式(5)中  $\alpha = 0.46$ ,结果如图 4 所示。由图 4 可知,引入广告主行为进行 Term 赋权,得到短语对相关性的准确率较高。因为单独使用 TF-IDF 是从 Term 独立的角度在全局计算 Term 在短语中表现的重要程度,但是相同的 Term 在不同文本环境下重要性会发生变化,往往不具有可比性。在广告语境下,更重要也更实用的是其内部 Term 之间的相对重要性,这个重要性因为和短语本身的构成和表达的语义有关,衡量其重要性也应该是一个动态的过程。ADPCB 采用全局赋权树动态获得短语中 Term 权值,解决了 TF-IDF 存在的上述问题,而且随着广告库数据集增大,应用将更为有效。

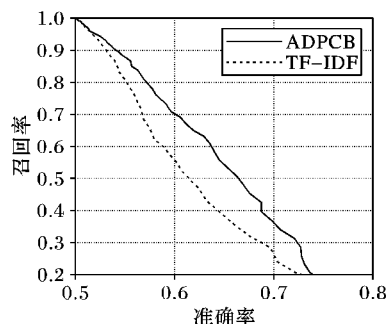


图4 引入广告主行为的效果评估

为了进一步验证算法的有效性,将 ADPCB 与常用的 LDA、BM25 进行对比,其中 LDA 计算前首先使用 2.1 节中的方法对 Query 和 Bidword 进行扩展,选取的主题数为 300,先验超参数取值:  $\alpha = 50/K, \beta = 0.01$ ; BM25 通常用于文档和 Query 相关性计算,为此首先将 Bidword 扩展为文本,式(1)参

数设置:  $K_1 = 2.0, K_2 = 2.0, avgdl = 378, b = 0.75$ 。图5结果表明ADPCB方法优于LDA和BM25。LDA和BM25算法为了适应短语相关计算需要进行短语扩展,目前的短语扩展方法存在较多冗余信息,而且部分扩展词可能导致语义漂移,影响了模型效果<sup>[15]</sup>。

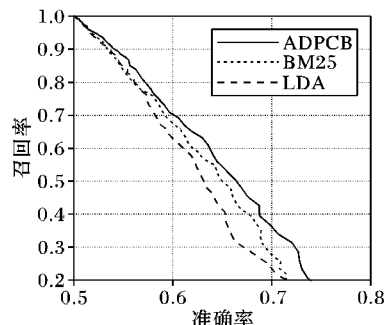


图5 ADPCB和其他模型对比

在拍卖词触发中,为了保证广告的相关性,需要严格控制准确率,在一定准确率的基础上要求召回率。因为  $revenue = CTR * Ad-show * CPC$ , 其中  $CPC$  (Cost Per Click) 为广告出价,  $Ad-show$  为广告展现量,所以广告展现越多收入就会越高。为了平衡收入和用户搜索满意度,将准确率固定为一个较高值(0.70)对比不同方法的召回率,结果如表1所示。

表1 准确率为0.70时不同方法的召回率对比

方法	召回率	F1
LDA	0.25	0.37
BM25	0.28	0.40
TF-IDF	0.28	0.40
ADPCB	0.36	0.48

#### 4.3 实时性分析

CBOW模型通过离线训练,线上可直接得到词的向量表示;并且使用广告主行为进行Term赋权,可以离线计算节点静态权值、边的权值、高频Query的Term权重然后组织成哈希表,将运算转化为查询。线上运算时,命中静态数据直接输出,否则构建全局赋权树,由子成分Term权值计算得到Query的Term权值。ADPCB方法具有较高的执行效率,可以应用于需要实时计算短语相关性的场景中。

## 5 结语

为了提高搜索引擎中广告的相关性,保证用户搜索满意度,本文充分利用广告主的行为信息结合CBOW模型,提出了广告语境下的短语相关计算方法ADPCB。实验中使用ADPCB计算Query和Bidword的相关值,在拍卖词过滤中表现出了较好的效果,可以应用于大规模数据的广告系统中。此外和以往方法相比,ADPCB通过分析广告主行为对短语Term动态赋权,权值可以表示短语的商业价值,用于控制非商业意图Query的广告触发数量<sup>[16]</sup>。如何有效引入更准确的商业意图或行业信息用于广告短语的相关性判断将是我们的下一步的研究工作。

#### 参考文献:

[1] ZHOU A, ZHOU M, GONG X. Computational advertising: a data-centric comprehensive Web application [J]. Chinese Journal of Computers, 2011, 34(10): 1805-1819. (周傲英, 周敏奇, 宫学庆. 计算广告: 以数据为核心的Web综合应用[J]. 计算机学报,

2011, 34(10): 1805-1819.)

- [2] CASTELLS P, FERNANDEZ M, VALLET D. An adaptation of the vector-space model for ontology-based information retrieval [J]. IEEE Transactions on Knowledge and Data Engineering, 2007, 19(2): 261-272.
- [3] ZHU Z, SUN J. Improved vocabulary semantic similarity calculation based on HowNet [J]. Journal of Computer Applications, 2013, 33(8): 2276-2279. (朱征宇, 孙俊华. 改进的基于《知网》的词汇语义相似度计算[J]. 计算机应用, 2013, 33(8): 2276-2279.)
- [4] LI B, LIU T, QIN B, et al. Chinese sentence similarity computing based on semantic dependency relationship analysis [J]. Application Research of Computers, 2004, 20(12): 15-17. (李彬, 刘挺, 秦兵, 等. 基于语义依存的汉语句子相似度计算[J]. 计算机应用研究, 2004, 20(12): 15-17.)
- [5] CARPINETO C, ROMANO G. A survey of automatic query expansion in information retrieval [J]. ACM Computing Surveys (CSUR), 2012, 44(1): 1-56.
- [6] HOFMANN T. Learning the similarity of documents: an information-geometric approach to document retrieval and categorization [C]// NIPS 1999: Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2000: 914-920.
- [7] HOFFMAN M D, BLEI D M, BACH F R. Online learning for latent Dirichlet allocation [C]// NIPS 2010: Proceedings of the 24th Annual Conference on Neural Information Processing Systems. Vancouver: NIPS, 2010: 856-864.
- [8] MIKOLOV T, CHEN K, CORRADO G S, et al. Efficient estimation of word representations in vector space [C]// Proceedings of the 2013 Workshop at ICLR, arXiv: 1301.3781. (2013-09-07) [2014-02-06]. <http://arxiv.org/abs/1301.3781>.
- [9] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality [C]// NIPS 2013: Proceedings of the 2013 Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2013: 3111-3119.
- [10] ROBERTSON S. Understanding inverse document frequency: on theoretical arguments for IDF [J]. Journal of Documentation, 2004, 60(5): 503-520.
- [11] HILLARD D, SCHROEDL S, MANAVOGLU E, et al. Improving ad relevance in sponsored search [C]// Proceedings of the third ACM International Conference on Web Search and Data Mining. New York: ACM, 2010: 361-370.
- [12] CHANG Y C, HILL M L. Query rewrite with auxiliary attributes in query processing operations: U. S. Patent Application 13/346,366 [P]. 2012-01-09.
- [13] MORIN F, BENGIO Y. Hierarchical probabilistic neural network language model [C]// Proceedings of the International Workshop on Artificial Intelligence and Statistics. Cambridge: Cambridge University Press, 2005: 246-252.
- [14] BENGIO Y, SCHWENK H, SENÉCAL J S, et al. Neural probabilistic language models [M]// Innovations in Machine Learning. Berlin: Springer-Verlag, 2006: 137-186.
- [15] GOTTIPATI S, JIANG J. Linking entities to a knowledge base with query expansion [C]// EMNLP '11: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2011: 804-813.
- [16] BRODER A, CIARAMITA M, FONTOURA M, et al. To swing or not to swing: learning when (not) to advertise [C]// Proceedings of the 17th ACM Conference on Information and Knowledge Management. New York: ACM, 2008: 1003-1012.