文章编号: 1000-5862(2012)04-0383-05

基于 Markov 网络团的查询意图识别

蔡桂秀, 王明文*, 揭安全, 王晓庆

(江西师范大学计算机信息工程学院,江西 南昌 330022)

摘要: 通过利用 Markov 网络团的方法来对查询意图识别. 首先从人工标注搜狗查询日志中约 2 250 个查询作为测试数据, 采用搜狗提供的分类语料(共 10 类)来建立 Markov 网络, 用建立的 Markov 网络来对查询进行扩展, 得到相关的返回结果列表, 运用在分类语料训练好的分类器来对返回结果进行分类, 从而完成对查询意图识别的过程. 实验中采用的评价指标是 11_avg 和 3_avg, 实验结果表明该方法能够有效地提高检索效率.

关键词: 查询分类; Markov 网络; 文本分类; 11_avg; 3_avg
中图分类号: TP 391.1　　文献标志码: A

0 引言

目前, 随着网络技术与计算机技术的日益发展, 互联网已经成为当前人们获取信息的主要来源之一. 根据中国互联网络发展状况统计报告中的数据显示, 互联网上的数据庞大, 并以指数级的方式增长, 并且互联网上的数据以文本形式为主.

搜索引擎如何为用户提供更好的信息服务, 查询意图分析是一种有效的组织和管理文本信息的工具, 它能发现大规模数据中潜在的有用模式. 如果能够理解隐藏在用户查询背后的意图, 就能够帮助搜索引擎自动将查询提交到相对应的垂直搜索引擎上, 从而得到更加相关准确的返回结果, 也就提高了用户的满意度. 比如新闻搜索、图片搜索、地图搜索、工作搜索等, 都提供一个特殊类别的信息服务, 以此来满足用户的特殊意图的需求. 但是将成千上万的查询准确地预测到某些特定的类别中去, 是一件困难的事情. 目前的大部分方法是采用统计机器学习方法来训练一个分类器. 根据统计机器学习理论, 为了对将来的查询有一个较好的泛化能力, 分类器需要 2 个条件：独立的特征表示和大量的训练样例. 然而, 对于意图分类的问题, 即使有大量的 Web 查询, 标注稀疏带来的问题也会使上述 2 个条件难以得到满足. 因此, 利用额外的知识来扩充查询能够较好地提高搜索质量. 另一方面, 更好地理解搜索意图能够让搜索引擎发现商业价值, 特别是搜索广告.

本文的主要工作包括：(1) 通过利用 Markov 网络来获得词之间的关系, 寻找每个词的团信息, 因为一个团中的词能够代表现实中一个概念. 在检索阶段根据与查询词在同一个团中的词对查询进行扩展, 从而得到更好的检索结果; (2) 通过对检索阶段得到的结果对查询进行意图判别, 在已知查询意图的情况下对查询进行再检索; (3) 实验对比了该方法与原始查询阶段的结果比较.

1 相关工作

查询意图识别是近年来信息检索领域的研究热点, 并且在很多领域得到了广泛的关注.查询的“短”问题至今没有得到很好地解决. 因此, 利用额外的知识来扩充查询能够较好地提高搜索质量. 另一方面, 更好地理解搜索意图能够让搜索引擎发现商业价值, 特别是搜索广告. Broder 等[1]在用户调查和对查询日志进行手工分类的基础上指出, 把查询意图分为导航类、信息类和事务类. 为了解决查询意图识别中的查询太短, 特征稀疏等问题, 人们相继提出了多种查询意图识别算法.

文献[2]提出了通过一个商业搜索引擎的广告点

收稿日期：2012-04-03
基金项目：国家自然科学基金(60963014)和江西省自然科学基金(20114BAB201037)资助项目.
作者简介：王明文(1965-), 男, 江西南康人, 教授, 博士, 主要从事信息检索和并行计算的研究.

击信息来学习用户的商业意图特征. 文献[3]介绍了每一个意图类别可以用一个维基百科中的概念或者目录集合来表示. 那么对于查询, 将其映射到这个表示空间中来进行分类. 这个方法对比于以前的方法, 一个很好的特性就是能够解决一些数据稀疏的问题. 文献[4]提出了 3 种独立的方法来进行查询主题的归类：人工标记的查询、有监督的分类、有限选择性, 最后将这 3 个方法组合起来. 文献[5]提出了一种更加有效地通过点击信息分布的查询意图分类的方法. 文献[6]指出点击信息已经很成功地用在了用户的查询意图推断中, 但是还是有很多的噪声和模糊性. 文献中开发了一种互补的而且更加敏感的特征：鼠标运动, 来识别导航类和信息类. 作者的假设是鼠标运动能够提供更多的有关用户交互的信息. 文献[7]提出了一种通过点击图来提高查询意图分类, 根据已有的少量点击图信息, 应用半监督的学习方法快速地增加训练数据信息.

Markov 网络在信息检索领域得到了广泛的应用, 文献[8]介绍了一种基于团的 Markov 网络查询扩展全局技术, 与以往工作不同点在于, 该技术首先提取文档集中的词频在一定范围内的所有词, 并计算词与词之间的相似度得到 Markov 网络, 对于每个查询, 使用团提取算法在网络中提取最相关的扩展词进行查询扩展, 取得了很好的检索效果. 文献[9]提出并实现了基于 Markov 网络的信息检索扩展模型, 通过对文档的学习, 构造了关于索引项和文档的 Markov 网络, 将有利的信息加入到检索中, 实验表明, 该模型优于 BM25 模型[10].

2 基于 Markov 团的查询意图识别

本文提出的基于 Markov 团的查询意图识别模型是由 2 个阶段组成的：初始查询阶段和加入查询意图后的查询阶段. 在图 1 中给出了 2 个阶段模型的 Markov 查询意图识别过程.

2.1 初始查询阶段

在初始查询阶段, 先对搜狗提供的分类语料和查询进行预处理; 得到词与词之间的关系网, 然后通过该网构造词之间的 Markov 团, 得到每个词的团信息. 对每个查询, 通过加入查询词的团信息对查询扩展, 从而计算文档与查询的相似度值. 通过 3_avg 和 11_avg 来度量检索的效果. 具体实现算法参照文献[8].

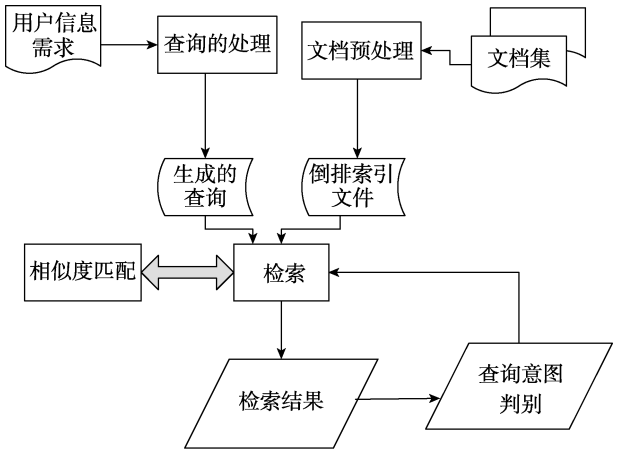

图 1 基于 **Markov** 团的查询意图识别流程图

2.2 查询意图判别

在意图识别阶段, 首先利用搜狗提供的分类语料的 10 个类别文档(每类 1 990 篇)进行预处理; 应用开源 Libsvm 训练分类器(该分类器也用于后面的查询分类), 查询分类的测试数据为从搜狗 2006 年 8 月的查询日志中选取并标注了约 2 250 个查询(每个查询仅属于一个意图).

在查询阶段中将待分类的查询 q 提交给建立好的 Markov 网络中获得查询对应的相关文档列表, 表示为 $D=\{d_1,d_2,\cdots,d_N\}$, 目标类别为 $C=\{c_1,c_2,\cdots,c_k\}$, N 表示查询对应的文档数, k 为意图数. 通过计算 $c^*=\arg\max\limits_{c_k}^{k=1,2,\cdots,K} p(c_k\,|\,q)$ 来确定查询的最终意图.

对每个意图 k 采用一种投票的方法来计算 $p(c_k\,|\,q)$. 对所有的文档列表采用训练得到的分类器对这些文档进行意图预测, 得到文档在各个意图上的概率 $P=\{P_1,P_2,\cdots,P_N\}$, 这里 P_i 为 K 维向量, $\sum\limits_{k=1}^{K} p_{i,k}=1$ 表示每个 P_i 在各个意图上的概率和为 1. 这里使用 $Score(c_k|q)$表示查询 q 属于意图 k 的分值：$Score(c_k\,|\,q)=\sum\limits_{i=1}^{N} P_{i,k}\big/\log(i+1)$, 也就是查询 q 对应的各个文档对意图 k 的所属概率加权求和. 根据文档在查询中的排序来确定权重系数 $1/\log(i+1)$, 即越靠前的文档权重也越大. 最后通过 $p(c_k\,|\,q)=Score(c_k\,|\,q)\Big/\sum\limits_{i=1}^{K} Score(c_i\,|\,q)$ 来确定 q 所属的意图信息.

2.3 优化查询

在优化查询阶段, 通过前面得到的查询意图,

把查询意图信息加入检索过程中, 并调整初始查询过程中的权重因子, 进行再次检索, 分析检索结果.

3　实验设计及结果分析

3.1　实验准备

本文使用搜狗查询日志及搜狗提供的分类语料进行实验. 该数据集总共有 10 大类别及约有 2 250 个查询. 去除一些不健康的查询和一些意图很明显的查询之后, 每类查询的个数见表 1.

在进行实验前, 首先需要对文档数据进行预处理, 步骤如下：(1) 去除文档中的格式标记、过滤非法字符、中文分词, 词频统计, 去除停用词等; (2) 生成文档矩阵, 查询矩阵, 相关矩阵等; (3) 特征过滤, 根据逆文档频率： $idf = \log(N/n_i)$ 来选择词; (4) 采用 LTC 权重公式计算文档中词项的权重.

对文档及查询数据, 均采用上述方法对它们进行预处理(文档和查询中有些参数需要调整), 以确保实验结果的可比性.

表 1　文本分类的训练数据及测试查询统计

类别	财经	IT	健康	体育	旅游	教育	招聘	文化	军事	汽车
样本数	1 400	1 400	1 400	1 400	1 400	1 400	1 400	1 400	1 400	1 400
测试查询数	100	145	108	110	120	126	80	170	93	148

3.2　实验设计

本文中主要进行了 2 次实验：一次是对选取的原始查询进行检索的实验; 另一次是引入查询意图之后的查询检索实验. 为了验证本文提出方法的有效性及可行性, 每次实验都进行 2 组实验：分别为 Markov 团(clique)的信息检索和 Markov 概念图(MCG_IR)的信息检索实验. 本文检索阶段使用的评价指标为 3_avg 和 11_avg.

3.3　实验结果及分析

为了验证方法的有效性, 本文使用了以下实验来进行对比：原始查询阶段的检索结果与引入查询意图后的检索结果如图 2 和图 3 所示.

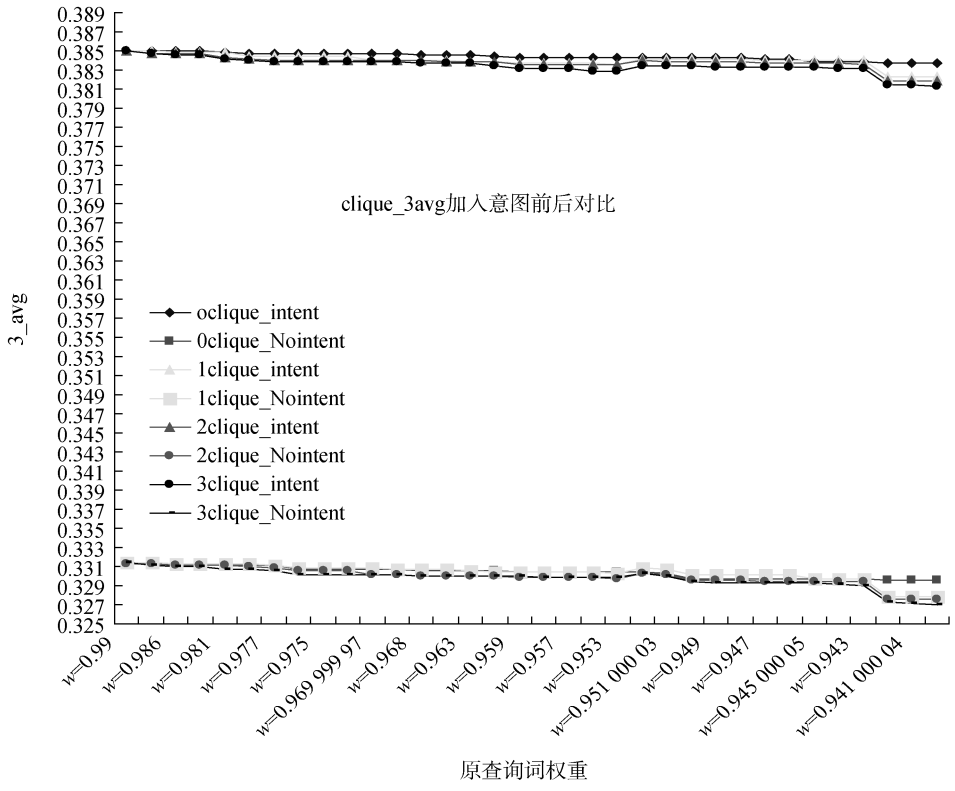

图 2　引入查询意图之后 **Markov** 团结构检索 **3_avg** 前后对比

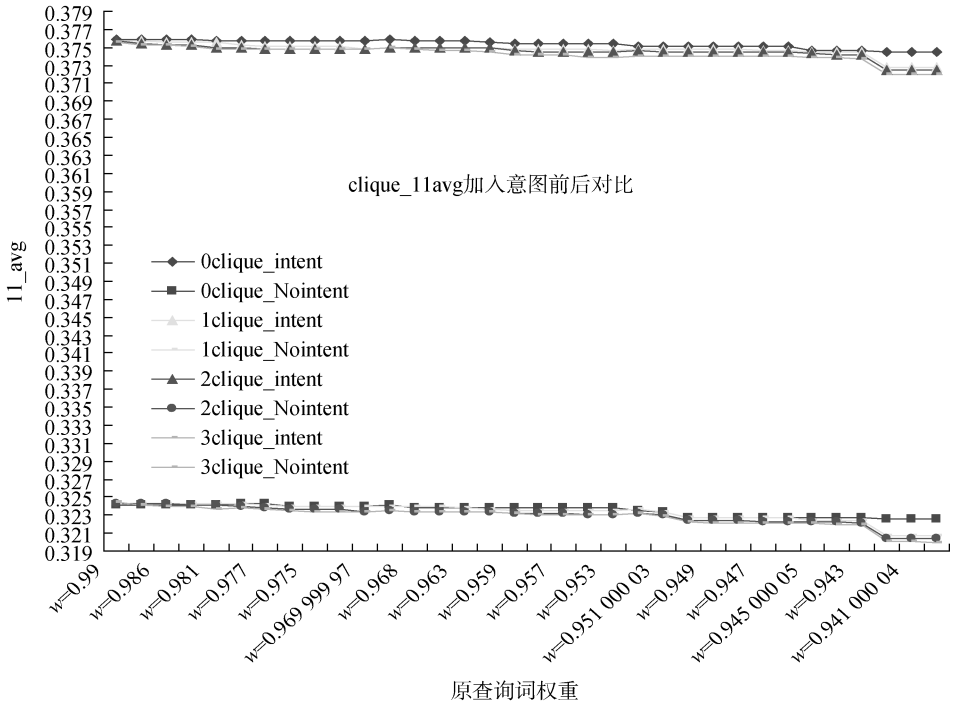

图 3　引入查询意图之后 Markov 团结构检索 11_avg 前后对比

在Markov团结构的实验过程中看到，加入检索意图之后，平均结果都能提高 16%左右. 因此在检索过程中加入检索意图是提高检索效率的有效方法之一. 在 MCG_IR 检索过程中平均提高达 21%左右，再次表明加入检索意图是提高检索效率的有效方法之一. 实验也表明了的团算法的优越性.

从以上实验结果可以发现，在检索过程中，加入检索意图之后的检索效率都有较大的提高，因为当加入检索意图之后，会有目的地去与查询意图相关的文档中检索，这使得检索效率以及检索精度都会有所增加.

4　结论

在信息检索中，作为一种有效的组织和管理信息的工具，查询意图识别已经被广泛的研究和应用，但随着信息科学与计算机技术的日益发展，在越来越多的应用中，人们对网上暴增的数据感到无所适

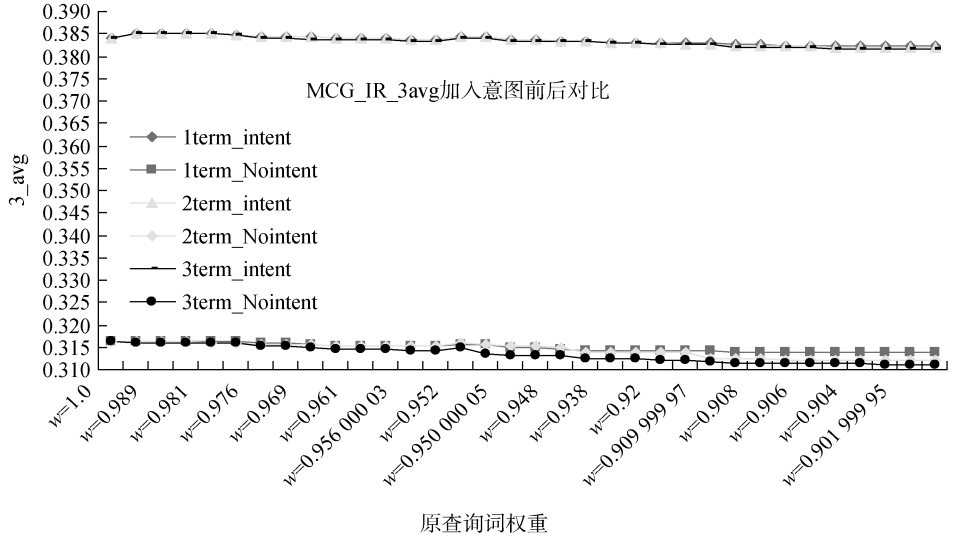

图 4　引入查询意图之后 MCG_IR 检索 3_avg 前后对比

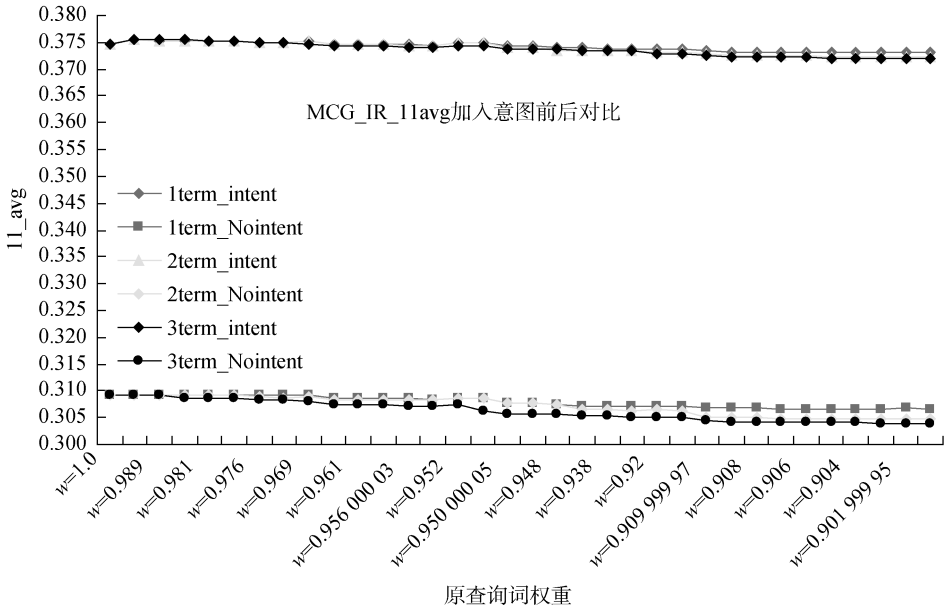

图 5　引入查询意图之后 MCG_IR 检索 11_avg 前后对比

从，因此，如果能够有效地对用户的意图进行预测，那么可以根据不同的策略提供给不同的结果．本文利用一种有效的提高检索准确率的方法来达到对用户意图的识别，这也正是本文研究的价值所在．

在信息检索过程中，由于查询信息往往很短，如果需要更好地对查询的意图进行分析，需要额外的信息来扩展查询，全局分析方法是查询扩展中一种常用而有效改善信息检索效果的查询扩展方法．本文提出的查询扩展的全局分析方法是利用词之间的团信息来对查询进行扩展，这种方法能够更有效地对查询进行扩展，并且通过扩展后的查询结果来对该查询进行意图识别，从而达到查询意图识别的目的，因此能够更好地满足用户的需求．

文本提出的查询意图识别具有实用和研究价值，未来的主要研究工作有以下几个方面：(1) 选择其他更好的度量词之间关系的方法；(2) 利用更大的文本数据集进行实验，进一步验证该方法的有效性；(3) 利用更好的分类方法对查询的结果进行分类．

5　参考文献

[1] 张森，王斌，张磊. Web 检索查询意图分类技术综述 [J]. 中文信息学报, 2008, 22(4): 75-82.

[2] Ashkan A, Clarke C, Agichtein E, et al. Characterizing query intent from sponsored search clickthrough data [EB/OL]. [2012-02-16]. http://www.mathcs.emory.edu/~qguo3/ira2008-azin.pdf.

[3] Hu Jian, Wang Gang, Lochovsky F, et al. Understanding user's query intent with Wikipedia [C]. New York: ACM, 2009: 471-480.

[4] Beitzel S, Jensen E, Frieder O, et al. Automatic web query classification using labeled and unlabeled training data [C]. New Youk: ACM, 2005: 581-582.

[5] Liu Yiqun, Zhang Min, Ru Liyun, et al. Automatic query type identification based on click through information [J]. Lecture Notes in Computer Science, 2006, 4182: 593-600.

[6] Guo Qi, Agichtein E. Exploring mouse movements for inferring query intent [EB/OL]. http://www.mathcs.emory.deu/~eugene papers/sigir2008p-mouse-moves.pdf.

[7] Xiao Li, Wang Yeyi, Alex Acero. Learning query intent from regularized click graphs. [EB/OL]. [2012-02-26]. http://research.Microsoft.com/pubs/75219/sigir2008.pdf.

[8] 甘丽新，王明文，张华伟. 基于团的 Markov 网络信息检索模型[EB/OL]. [2012-03-19]. http://www.docin.com/p-193591277.tml..

[9] 左佳莉，王明文，王希. 基于 Markov 网络的信息检索扩展模型 [J]. 清华大学学报: 自然科学版, 2005, 45: 1847-1852.

[10] Makoto Iwayama, Astushi Fujii, Noriko Kando, et al. An empirical study on retrieval models for different document genres: patents and newspaper articles[C]. New York: ACM, 2003: 251-258.

(下转第 394 页)