机器学习的查询扩展在博客检索中的应用

王秉卿， 张 奇， 吴立德， 黄萱菁
(复旦大学 计算机科学与工程系，上海市 200433)
通讯作者：王秉卿，电话：+86-21-65642830-315,电子邮件：wbq@fudan.edu.cn

摘 要：本文介绍一种新的查询扩展方法。该方法将查询扩展工作纳入机器学习的框架下，首先伪反馈将生成原始查询项的候选扩展词集合，然后一个支持向量机将对这些候选词进行排序，形成一个优化的查询项，以此来提高最终检索结果的性能。由于此类方法所需的训练数据较难获得，文中还介绍了一种新的自动生成训练数据方法。本方法的优点在于通过对训练语料的学习，能够对候选扩展词作出更合理的选择。通过在 TREC 的 BLOG TRACK 的观点检索任务的检验，此方法取得了良好的结果。

关键词：信息检索；查询扩展；机器学习；

Machine Learning based Query Expansion in Blog Retreival

Wang Bingqing, Zhang Qi, Wu Lide, Huang Xuanjing
(Department of Computer Science, Fudan University, Shanghai 200433, China)
+ Corresponding author: Wang Bingqing Phn: +86-21-65642830-315, E-mail: wbq@fudan.edu.cn

Abstract: This paper is aimed to introduce a new query expansion approach, which adopts the machine learning technique. Pseudo-relevance feedback would first generate a set of candidate expansion words for a certain topic. Then a Support Vector Machine (SVM) would predict on these candidate words and rank these words. A better query could be formed with the top candidate words to achieve better retrieval performance. It is difficult to get reliable training data to train the SVM, a new approach was proposed to generate the training data set. By learning from the training data, the SVM could rank the candidate words more reasonably, which is the advantage of this query expansion approach. This approach was applied in the opinion retrieval task of BLOG TRACK held by the TREC conference and got good experiment results.

Keywords: Information Retrieval; Query Expansion; Machine Learning;

1. 前言

在文本检索中，由于待查询的话题提供的查询项比较简短，可提供的信息不充分，使得最终的检索结果性能不佳。查询扩展方法寻找与待查询的话题相关的词或短语作为扩展词，扩展词的引入可有效地提高系统性能，尤其对于大规模语料，查询扩展方法能较好地提高检索结果的各项评测指标[1,2]。优化查询项可以有效提高系统最终的检索结果，因此需要挑选合适的扩展词来优化查询项。

对于查询扩展方法的研究已有大量的工作。早期的研究通过手工或自动的方式构建同义词近义词词典来生成扩展词[5,6]；随着 WordNet，Wikipedia，Google 等资源的流行，研究工作也关注如何利用知识库或搜索引擎来生成扩展词[3,7]。除了从外部资源挖掘，人们也研究从返回的文

档集中分析获取扩展词，Rocchio 早在 1970 年在向量模型下提出了查询扩展的框架，这套框架在 Salton 的 SMART 系统中采用。Xu 和 Croft 综合局部返回的文档集合和全局的文档集合对候选词进行评分[4]。Cappineto 和 Romando 基于信息论在 Rocchio 的框架下改进了候选扩展词的评分方法[2]。这种从返回的前若干篇文档中抽取扩展词的方法，称为伪反馈（pseudo-relevance feedback 或者 local feedback）。伪反馈实现简单，但对候选词的评价基于经验公式上，对一些候选扩展词的评分不可靠。

另一方面，一些研究运用机器学习的方法来提高检索结果。检索输出的文档集合会有相应的评测结果，在这些评测结果上，以提高检索结果的评测指标为目标，训练一个机器学习模块，这个模块对于返回的文档进行重新排序[12,13,14]，以此来提高系统检索结果的性能。这种以学习的方式进行排序（Learning to Rank）是近年来的一个研究热点。

鉴于上述的背景，我们希望将查询扩展的方法和机器学习的理论结合起来，通过机器学习训练出来的模块来对候选词打分排序，优化输入的查询项，并且优化最终输出的检索结果。以学习的方式对候选扩展词进行排序，可以使生成的扩展词更合理可靠。对扩展词排序和对文档进行排序都是将学习的方法应用到检索中，扩展词可以提取出丰富的特征，并且实现简单。这套方法的难点在于难以获得训练数据，我们提出了一种新的自动生成训练数据的方法。在 TREC 评测会议组织的博客检索（BLOG TRACK）的观点检索任务中[8,9]，我们采用了这套查询扩展方法并取得了良好的实验结果。

本文的结构如下：第 2 部分包含本方法的实现过程，第 3 部分介绍训练数据自动生成的方法，第 4 部分介绍本方法在 TREC 的观点检索任务中的应用，第 5 部分介绍实验结果，第 6 部分是总结。

2. 查询扩展和机器学习

基于机器学习的查询扩展方法，其实现的流程如图-1 所示。对于需要检索的语料库建立全文索引后，输入的查询项通过全文索引返回初始查询结果，由伪反馈方法从返回的文档集合中生成候选扩展词。经过训练得到的支持向量机（SVM）对候选扩展词进行评分，挑选其中得分较高的扩展词和输入的查询项组成优化的查询项，并通过全文索引得到输出的检索结果。

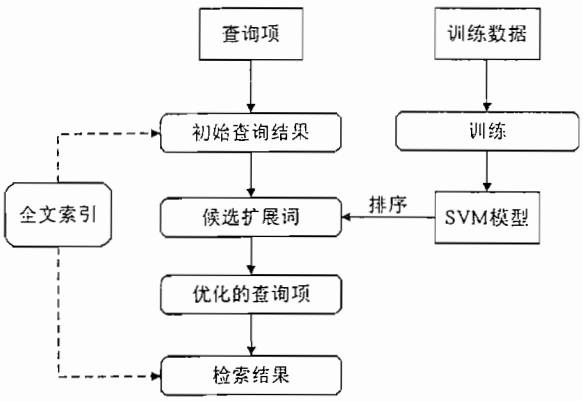

图-1 查询扩展流程

2.1 候选扩展词生成

我们的检索系统采用向量模型，由伪反馈方法生成候选扩展词集合。标准 Rochio 公式[15]给出了向量模型下的查询扩展框架，对 Rochio 公式演化可以有如下的形式[2]。

$$\vec{q}_m = \vec{q} + \lambda \sum_{\forall \vec{d}_j \in D_r} \vec{d}_j = \vec{q} + \lambda \sum_{\forall \vec{d}_j \in D_r} (w_{1,j}, w_{2,j} \cdots w_{n,j}) = \vec{q} + \lambda (w_1, w_2 \cdots w_n)$$

$$w_{i,j} = tf_{i,j} \cdot \frac{idf_i}{sidf_i} \qquad w_i = \sum_{\forall \vec{d}_j \in D_r} w_{i,j}$$

其中，$\vec{q}_m$ 表示新的查询项，$\vec{q}$ 表示原始查询项，D_r 是初次查询后系统输出的前若干篇文档组成的文档集合。$\vec{d}_j$ 代表一篇文档的权重向量，λ 是归一化系数。

我们采用一种新的公式计算权重 $w_{i,j}$，$w_{i,j}$ 是第 i 个词在第 j 篇文档中的权重，w_i 是第 i 个词在 D_r 上的权重。$tf_{i,j}$ 是第 i 个词在第 j 篇文档中的词频，idf_i 是第 i 个词在整个语料库上的反文档比，$sidf_i$ 是第 i 个词在 D_r 上的反文档比。词的反文档比可近似代表这个词在文档集合中的分布，当一个词在整个语料库上的分布和在 D_r 上的分布近似时，这个词的得分会降低。

λ 是归一化系数，$\lambda = \dfrac{0.5}{\max(w_i)}$，这个系数使得任何扩展出来的词的权重小于 0.5，采用这样的归一化公式是假定任何扩展词的权重都不应该高于原始的查询项。

根据 w_i 计算得到各个词的权重，取其中得分较高的词作为候选扩展词集合。

2.2 支持向量机和扩展词排序

对于大规模语料库，一个查询项可以抽取出许多的候选扩展词，不相关的扩展词作为噪音会产生话题漂移，影响检索结果，因此需要对扩展词进行挑选。

扩展词可以抽取出许多统计特征，例如总词频数，平均词频数，反文档比，互信息等。采用经验公式对候选扩展词进行打分，只是利用其中一部分统计信息。要设计出一套经验公式去充分利用所有这些信息是比较困难的。

支持向量机[11]采用有监督的机器学习方法。作为一种线性分类器，应用于分类问题和回归问题。在学习阶段，通过学习训练数据集合：$(\vec{x}_i, y_i)$, $i = 1, 2 \cdots, N$，获得一个支持向量机模块。在测试阶段，用这个训练出的模块对于输入的测试数据进行分类，回归等操作。支持向量机的在学习阶段的工作是解决如下的优化问题：

$$\min \ \frac{1}{2} \| w \|^2 + C(\sum_i \xi_i)^k$$

$$s.t. \quad y_i(w^T x_i + b) - (1 - \xi_i) \geq 0,$$

$$\xi_i \geq 0, \quad i = 1, 2, \cdots, N$$

将支持向量机应用到扩展词排序中。一个训练样本 $(\vec{x}_i, y_i)$ 对应一个扩展词，$\vec{x}_i$ 表示这个扩展词的特征向量，y_i 表示这个扩展词的得分。在学习阶段，将这个任务看作是回归问题进行学习。在测试阶段，将候选扩展词的统计特征作为输入训练得到的支持向量机，支持向量机输出

这个词的得分，根据这个得分对于候选扩展词进行排序。

由于一个扩展词的所有统计信息均可以作为特征，因此采用机器学习的方法可以更充分的利用候选扩展词的信息，对候选扩展词的评分可以更可靠。

3. 训练数据的生成

训练数据的生成是本方法的一个难点，也是这套方法的一个贡献。用于训练候选扩展词得分的训练数据难以获得。对于训练数据集合$(\vec{x}_i, y_i)$, $i = 1,2\cdots,N$，一个扩展词的特征向量$\vec{x}_i$可以获得，但是这个扩展词的得分y_i难以生成。人工标注费时费力，更困难的是人工标注不准确，标注人员无法深入地了解每一个输入的查询项，无法判别一个候选扩展词是否和这个话题相关，更无法准确的给这个候选词评分、排序。例如，输入的查询项是“March of the Penguins”，若候选扩展词是 morgan，标注人员很难判断这个词是否和话题相关，而实际上 morgan 是 “March of the Penguins”这部纪录片的配音，是一个良好的扩展词。

对于检索系统输出的结果，通常有评测工具评价检索结果的性能，例如可以评测检索结果的 MAP，R-Prec，b-Pref 等指标。对于一个候选扩展词的评分可以看作是这个候选扩展词能够在多大程度上提高检索结果的性能，得分越高说明这个词越能帮助提高系统的评测指标。基于这样的想法，充分利用过去的评测结果和评测工具来自动生成训练数据。

> 训练数据自动生成方法：
>
> 1） 输入查询项 q，输出检索结果，对该检索结果评测，令对应的 MAP 指标为 BaseMAP
> 2） 抽取查询项 q 的候选扩展词集合（$t_1, t_2, \cdots$）
> 3） 对每个候选扩展词 t_i
>
> a) 计算特征向量 $\vec{x}_i$
>
> b) 新的查询项 q_i←原始查询 q+ 候选词 t_i
> 输入 q_i，输出检索结果，对其评测，
> 令对应的 MAP 指标为 $\mathrm{TermMAP}_i$
>
> c) $y_i = \frac{TermMAP_i}{BaseMAP}$，（$\vec{x}_i$， y_i）是候选词 t_i对应的训练数据。

扩展候选词的特征向量$\vec{x}_i$，主要是这个扩展词的统计特征。用y_i来测量这个候选扩展词对于提高 MAP 指标有多大帮助，同时也可以表示这个扩展词和原始话题之间的相关性。$y_i > 1$表明这个候选扩展词有助于提高 MAP，$y_i < 1$表明这个词不能提高 MAP。

4. 观点检索中的查询扩展

TREC 评测会议组织的博客检索（BLOG TRACK）旨在探索对博客语料的特性及对这类语料进行处理的技术[8,9]。其中，观点检索任务要寻找针对某话题进行评论的文章，而不仅仅是对于这个话题介绍或者阐述性的文章。这项任务需要结合文本检索技术和主观性倾向分析技术。

TREC Blog06 Test Collection[10]作为 BLOG TRACK 2006, 2007 两年的测试数据，该测试数据集从 2005 年 12 月至 2006 年 2 月从互联网上共收集了 3215171 篇博客文章，数据解压后 148GB。

对于观点检索的任务，如图-2 所示，系统由预处理、话题检索和情感倾向分析组成。由于语料是 HTML 页面文件，预处理部分从语料中抽取文档的主要内容。话题检索的功能是对于输入的查询项进行检索，返回文档集合供情感倾向分析。在这一部分，我们采用了基于机器学习的查询扩展方法来优化查询项。在这个部分输出的文档集合，每篇文档会给予一个相似度（Similarity）表示这篇文档和话题的相关性。情感倾向分析部分，采用了基于句子级的情感倾向分析方法，训练一个 CME 分类器对每句句子的情感倾向性评分，综合一篇文章中所有的句子，对这篇文章给出一个情感倾向分数（Opinion Score）。最后，综合话题检索得到的相似度（Similarity）和情感倾向分分数（Opinion Score），对返回的文章重新排序。

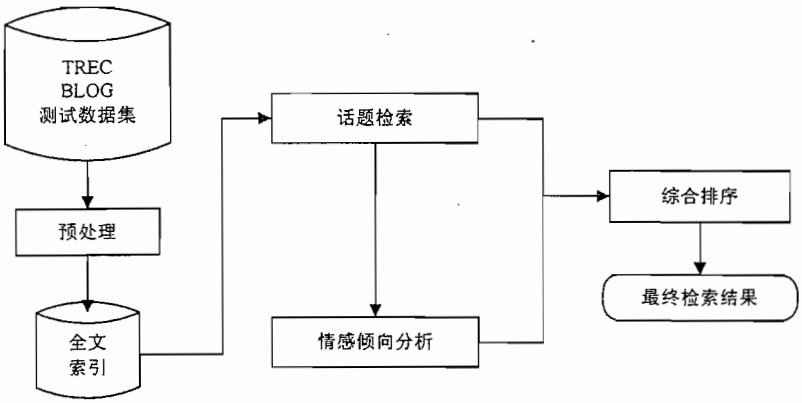

图-2 系统框架

观点检索任务给定 50 个话题（Topic），每个话题仅使用标题（Title）部分作为查询项，要求每个话题可返回至多 1000 篇文档作为检索结果。根据这样的要求，在话题检索部分，根据输入的查询项，首先取前 100 篇文档用来抽取候选扩展词集合，用训练好的支持向量机对这些候选扩展词评分，取得分高的候选扩展词与原始查询项一起组成新的查询项，通过检索系统输出前 2000 篇文档，再由情感倾向分析模块进行处理。

5. 实验结果

SVM light 被用来训练一个 RBF 回归模型，在 BLOG 06 的检索结果上生成训练数据进行学习训练，在 BLOG 07 上进行测试。BLOG 06 提供了两种评测，观点检索评测是和话题检索评测。观点评测仅把那些对话题进行评论的文章作为相关文档，而话题相关性评测把所有介绍或讨论这个话题的文章均作为相关文档看待。因此相应的，我们训练了两套支持向量机回归模型。用观点评测训练出来的模型有一定的情感倾向分析的特性。

如下表所示，在 BLOG 06 上进行训练，在 BLOG 07 上进行测试。共有 7 个结果

Baseline：不做任何查询扩展处理得到的基准结果

Run 1：以经验公式对候选词排序得到的结果

Run 2：机器学习的查询扩展方法，用 BLOG 06 年上话题相关性评测训练生成的支持向量机

Run 3：机器学习的查询扩展方法，用 BLOG 06 年上观点检索评测训练生成的支持向量机

Run 4：在 Run 1 的基础上进行情感倾向分析

Run 5：在 Run 2 的基础上进行情感倾向分析

Run 6：在 Run 3 的基础上进行情感倾向分析

表-1 观点检索评测

Run	MAP	R-prec	b-Pref	P@10
Baseline	0.2388	0.3011	0.3083	0.3680
Run 1	0.2992	0.3351	0.3357	0.4340
Run 2	0.3178	0.3447	0.3498	0.4520
Run 3	**0.3179**	**0.3467**	**0.3501**	**0.4540**
Run 4	0.3019	0.3382	0.3381	0.4460
Run 5	0.3141	0.3475	0.3496	0.4620
Run 6	0.3143	0.3465	0.3499	0.4600

表-2 话题相关性评测

Run	MAP	R-prec	b-Pref	P@10
Baseline	0.3927	0.4520	0.5222	0.6340
Run 1	0.4506	0.4744	0.5272	0.6320
Run 2	0.4709	0.4888	0.5428	0.6520
Run 3	**0.4714**	**0.4889**	**0.5432**	**0.6540**
Run 4	0.4355	0.4626	0.5113	0.6500
Run 5	0.4484	0.4765	0.5228	0.6620
Run 6	0.4488	0.4768	0.5232	0.6620

对于观点检索评测，当采用经验公式进行查询扩展时， MAP 值上升了 25.5%，当采用机器学习的查询扩展方法时，MAP 值相对 Baseline 上升了 33.12%；对于话题相关性评测，采用经验公式进行查询扩展时，MAP 值相对于 Baseline 上升了 14.7%，当采用机器学习的查询扩展方法时，MAP 值相对 Baseline 上升了 20.0%。

从结果中可以看出，采用机器学习的查询扩展方法可以得到更好的效果。而句子级的情感倾向分析存在很大的挑战性。若仅使用经验公式对候选扩展词进行排序，则情感倾向分析能一定程度上提高观点检索的性能，而采用了基于机器学习的查询扩展方法后，情感倾向分析对系统性能的提高帮助不大了。

6. 结论

在这篇文章中，我们介绍了基于机器学习的查询扩展方法，该方法在博客检索的观点检索任务中取得了良好的实验结果。基于机器学习的查询扩展方法难点在于训练数据的生成，我们提供了一种从评测结果中自动生成训练数据的方法。在未来的工作中，我们希望能更深入地研究如何用机器学习的方法来提高检索的性能以及情感倾向分析的方法。

参考文献

[1] Ed Greengrass, "Information Retrieval: A Survey", 30 November 2000

[2] Claudio Carpineto, Renato De Mori, Giovanni Romano, Brigitte Bigi "An Information-Theoretic Approach to Automatic Query Expansion", ACM Transactions on Information Systems, Vol. 19, No. 1, January 2001, Pages 1–27.

[3] Claudio Carpineto. and Giovanni Romano,. "Effective reformulation of Boolean queries with concept lattices." In FQAS 98, Roskilde, Denmark. Springer-Verlag, Heidelberg, Germany, 83–94.

[4] Jinxi Xu and W. Bruce Croft, "Query Expansion Using Local and Global Document Analysis", In SIGIR '96, Zurich, Switzerland, Aug. 18–22, ACM Press, New York, NY, 4–11.

[5] Brajnik, G., Mizzaro, S., and Tasso, C. "Evaluating user interfaces to information retrieval systems: A case study on user support." In SIGIR '96, Zurich, Switzerland, Aug. 18–22, ACM Press, New York, NY, 128–136.

[6] Cooper, J. W. and Byrd, R. J. 1997. "Lexical navigation: Visually prompted query expansionand refinement". DL '97, Philadelphia, PA, July 23–26, R. B. Allen and E. Rasmussen, Chairs. ACM Press, New York, NY, 237–246

[7] Rila Mandala, Takenobu Tokunaga, and Hozumi Tanaka, "Combining Multiple Evidence from Different Types of Thesaurus for Query Expansion", In SIGIR '99, Berkeley, California, USA, ACM Press, New York, NY, 191–197.

[8] Iadh Ounis, Maarten de Rijke, Craig Macdonald, Gilad Mishne, Ian Soboroff. "Overview of the TREC-2006 Blog Track", In TREC 2006

[9] Craig Macdonald, Iadh Ounis, Ian Soborof, "Overview of the TREC2007 Blog Track", In TREC 2007

[10] Craig Macdonald and Iadh Ounis. "The TREC Blog06 Collection: Creating and Analysing a Blog Test Collection" DCS Technical Report TR-2006-224. Department of Computing Science, University of Glasgow. 2006.

[11] Christopher J.C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition", In Data Mining and Knowledge Discovery, 2, 121–167 (1998)

[12] Yisong Yue, Thomas Finley, Filip Radlinski, Thorsten Joachims, "A Support Vector Method for Optimizing Average Precision", SIGIR'07, July 23–27, 2007, Amsterdam, The Netherlands.

[13] Thorsten Joachims, "A Support Vector Method for Multivariate Performance Measures", in Proceedings of the 22 nd International Conference on Machine Learning, Bonn, Germany, 2005.

[14] Thorsten Joachims, "Making Large-Scale SVM Learning Practical", 'Advances in Kernel Methods - Support Vector Learning', Bernhard Scholkopf, Christopher J. C. Burges, and Alexander J. Smola (eds.) MIT Press, Cambridge, USA, 1998.

[15] Ricardo Baeza-Yates, Berthier Ribeiro-Neto "Modern Information Retrieval" pp. 117-118