面向对话语料的标签推荐

房冠南[1]，袁彩霞[1]，王小捷[1]，李 江[2]，宋占江[2]
[1]北京邮电大学 计算机学院 智能科学与技术中心，北京 100876
[2]诺基亚北京研究院，北京 100176
E-mail: gnfang@gmail.com; {yuancx, xjwang}@bupt.edu.cn
li.jiang84@gmail.com; {zhanjiang.song}@nokia.com

摘 要：本文提出了一种针对对话语料的自动标签推荐方法——KeyEx。该方法首先基于加权 TFIDF 进行关键词抽取，加权因子融入对话者权重、句子重要程度和句子长度等因素；然后，通过频繁模式匹配进行关键词的二元扩展获取信息含量大的二元关键词；最后在同一尺度下对候选关键词进行排序得到 top-n 推荐标签。在 10,265 段中文对话上的实验结果表明，KeyEx 优于已有的 KNN 和 TextRank 模型。

关键词：标签推荐；中文对话；关键词抽取；对话因子；二元扩展

Tag Recommendation for Dialogue Corpus

Fang Guannan[1], Yuan Caixia[1], Wang Xiaojie[1], Li Jiang[2], Song Zhanjiang[2]
[1] Center for Intelligence Science and Technology, School of Computer, Beijing University of Posts and Telecommunications, Beijing 100876
[2] Nokia Research Center, Beijing 100176
E-mail: gnfang@gmail.com; {yuancx, xjwang}@bupt.edu.cn
li.jiang84@gmail.com; {zhanjiang.song}@nokia.com

Abstract: This paper introduces a novel method, named KeyEx, designed for automatically recommending tags for Chinese dialogues. The method first extracts keywords from the dialogue using TFIDF weighting and dialogue factors, such as speaker weight, sentence salience and sentence length. Then, it uses a bigram expansion module to extract informative bigram keywords through frequent pattern matching and finally it ranks the candidate tags under a uniform metric and selects the top ranked as recommended tags. Experiment results on 10,265 Chinese dialogues show that KeyEx outperforms previous models like KNN and TextRank.

Keywords: tag recommendation; Chinese dialogue; keyword extraction; dialogue factor; bigram expansion

1 前言

随着网络信息的迅猛增长，人们希望海量文本能被标记上合适的词标签，即用一个或几个词对文本内容进行描述，这样可以极大地加快人们的浏览速度。同时，好的标签对于提高文本分类、信息检索等自然语言处理任务的性能也具有极大的帮助。因此，出现了不少自动生成文本标签（标签推荐）的研究。

早期的标签推荐技术研究主要集中在对网络资源打标签，如网页文本、照片等。基于协作过滤的标签推荐是目前研究较广泛的方法(Marinho, et al., 2008; Graham et al., 2008; Shepitsen et al., 2008)。该方法为待标资源在训练集(已知标签的资源)中寻找相近的资源，将相近资源的标签进行加权排序后推荐给待标资源作为其标签。这种方法依赖具有标签的训练语料和标签集。而实际上，很多资源缺乏标签信息，中文资源尤其严重，使得该方法存在很大的局限性。

基于文本分析技术，从相似文本内容抽取关键词作为标签是另一类重要的标签推荐技术。Oliveira et al. (2009)构建了基于文本内容的标签推荐系统 Tess，主要是运用余弦相似度测量找到和已知文档最相近的文档集获取有效描述已知文档的所有词，然后通过词频排序得到推荐标签。该方法依赖

文档关系的稀疏程度，若文档间关联很小，推荐标签质量便会非常差，而且如果文档短小而不足以提供自身有效描述词，也会进一步恶化标签质量。

Krestel et al. (2009)提出基于 Latent Dirichlet Allocation (LDA)的标签推荐，他们使用已标注的资源和一个相当稳定、完整的标签集合来为一些仅有少量标签的资源抽取潜在主题。虽然，这种方法可以产生具有普通视角的标签，但前提是资源必须已有少量标签。

近年来，即时聊天、twitter、微博等各种方式的社会化网络迅猛增长，人们基于这些工具表达和交流他们的观点，这类数据与已有的网页文本有很大的差异，比如都具有一定的对话的特点、通常文本较短、结构松散(经常会有多人参与从而成为多方对话)。这些特点使得其标签的遴选存在更多的困难。目前直接针对这类文本的标签推荐研究还很少见，针对一般网页文本的标签推荐方法能否在这类数据上具有良好的性能仍然未知。

本文关注于这类具有多方对话特性的短文本，提出了一种自动标签推荐方法。该方法首先利用融合了多种加权因子的 TFIDF 框架进行关键词抽取；再根据预先定义的 POS 模板序列进行过滤得到高频二元关键词；最终，在同一度量标准下，从关键词和二元关键词中推荐 top-n 的标签。我们采用两种不同的度量标准来评估该方法。一个是自动评价，这种评价方式对比人工标签与系统推荐标签，并在 Top-k 准确度，Exact-k 准确度，精确度-召回率@topN 下评价系统性能。另外一个是使用拒绝率进行人工评价。实验表明，与基于协同过滤的最近邻(KNN)标签推荐和基于图的 TextRank 模型相比，KeyEx 具有更好的性能。

2 KeyEx 方法

本文的标签推荐主要分两步。首先利用融合了多种加权因子的 TFIDF 框架进行关键词抽取；然后对关键词进行二元拓展，并在统一尺度下进行排序，得到 top-n 个标签。

2.1 标签抽取

在对语料进行关键词抽取之前，针对对话语料的特点，在停用词词表中增加了不包含主题信息的打招呼（如“你好”）、一般性应答（如“谢谢”）用语。我们考虑如下几个影响因素：

（A）TFIDF 基础加权

在文档中，词频（TF）指该单词在文档中出现的次数。逆文档频率 IDF 的值为：log(N/Ni)，Ni 指在文档集合中，包含该词的文档个数，N 指在文档集合中该词出现的总次数。

（B）POS 过滤

Hulth 在(Hulth, 2003)中指出，动词、名词和形容词在文档中表达了最重要的观点。基于此，我们利用这些 POS 标注来限制我们对候选关键词的选择。

（C）对话者信息

对话者信息是对话文档的特有属性，我们使用对话者权重来衡量对话者在文档中的重要程度，本文主要考虑了三种衡量方式。

Ⅰ、对话者说出的句子数占整个对话包含句子的比重，如公式(1)；

$$S\_spr\_weight = SentenceNum(spr) / SentenceNum(C) \quad (1)$$

Ⅱ、对话者说出的词数占整个对话包含词数的比重，如公式(2)；

$$W\_spr\_weight = WordNum(spr) / WordNum(C) \quad (2)$$

Ⅲ、对话者说出的实词数（主要考虑动词、名词和形容词）占整个对话包含的实词数的比重，如公式(3)；

$$CW\_spr\_weight = ContentWordNum(spr) / ContentWordNum(C) \quad (3)$$

其中，S_spr_weight、W_spr_weight 和 CW_spr_weight 均代表对话者权重，spr 代表对话者，C

代表整个对话。

（D）句子重要程度信息

TFIDF 的一个缺点是只考虑了词频，没有考虑对话的结构信息。为了平衡句子信息，通过包含候选词的句子重要性调整候选词权重。这里，句子重要性权值通过句子向量模型在整个对话空间的相似度计算。如果一个候选词出现在几个句子中，我们将使用这些句子的最高权值。这个权值通常用在摘要抽取中(Radev et al., 2001)使用。一个候选关键词的句子重要性的定义如下：

$$SW(w_i) = Max\_Sim(S_i, C) \tag{4}$$

其中，S_i 代表包含词 w_i 的空间向量，C 代表整个对话的空间向量。

（E）句子长度信息

直观上，越长的句子包含越多的信息。通过计算包含候选关键词的句子中包含的所有词个数衡量句子长度。若一个候选词出现在几个句子中，我们将使用这些句子的最大长度。同一个句子中的候选关键词，具有相同的句子长度。句子长度定义如下：

$$Len(w_i) = Max\_WordNum(l) \tag{5}$$

其中，w_i 代表候选关键词，l 代表 wi 所在的句子。

2.2 二元词扩展

通过分析人工标注集，我们发现二元关键词在人工标注标签中占了大约 22%，包含三个词及以上的标签几乎不存在。所以，抽取二元词是有必要的。我们使用了 75 种不同的 POS 标注（词性标注集（ZHANG et al., 2003）），其中存在 65 种相邻词的 POS 组合。在初步的实验语料中发现其中 18%的 POS 组合在发展集中占 89%。因此，我们可以通过 POS 模板来获取二元关键词，二元关键词是基于上述加权 TFIDF 抽取的原始关键词的拓展。表 1 中展示了几个二元 POS 模板。二元关键词的权值由其包含的一元关键词的平均权值确定而不使用二元关键词在文档集中的频率信息来计算其权值，实际上，二元标签在文档集合中出现的频率均低于 3 次。其计算公式为(6)。

表 1　POS 组合的例子

a+a	ad+v
a+an	an+n
a+f	an+vn
a+n	v+n
a+vn	v+v

$$W(biWord) = \frac{W(keyWord1) + W(keyword2)}{2} \tag{6}$$

其中，biWord 是二元候选关键词，keyWord1 和 keyword2 是其包含的一元关键词。二元关键词的权值由其包含的关键词决定，同时也保证了基于加权 TFIDF 抽取的候选标签和二元扩展后的候选标签能够在同一尺度下进行排序，最终得到权值在 top-n 的标签，作为推荐标签。

3 实验

本节介绍相关比较实验。

3.1 实验设置

实验使用的数据集是来自访谈节目的对话，我们采用相关主题切分技术，将对话文档切分成了对话段落。本文以下的“对话”均指具有同一主题的“对话段落”。

我们招募了 10 名计算机学院的学生进行人工标注，每两个学生组成一组标注相同的对话，总

共标注了 10,265 个对话。对话的平均长度（词的个数）是 148，每个对话的平均标签个数是 3.2。我们使用了 950 个对话作为发展集，5,114 个对话作为对比方法（4.3 节）的训练集，剩余的 4,201 个对话作为最终的测试集。

人工标注例子如下：

- 标注者 1：对话，经济，改革开放
- 标注者 2：economic，邓小平，繁荣

这个例子针对的对话可能是关于经济的，由于标注者看待该对话的角度不同，对话中的句子或者标签的重要程度对于不同标注者是不同的。另外，由于标签并未限制在特定的词汇表中，标注者可以挑选任意他们喜欢的标签来描述资源，这将会导致标签的不连贯性或特殊性。

为了更好的了解该任务中的黄金评价标准，我们分析了人工标注的一致性。我们使用 3 组人工标注的 950 个对话来计算两个不同标注者间的一致性。所有的一致性分析都是基于严格匹配的。两个标注者对于同一对话的一致率是 30.14%，方差高达 70.5。

3.2 实验结果

我们首先通过实验分析基于加权 TFIDF 进行关键词抽取时各种加权因素的贡献，以选出能较好提高系统性能的因子组合。我们在特征因子集合上做了一系列的对比实验。在每次实验中，通过增加一个特征因子，利用增量-驱动的方式检查该特征因子是应该保留还是删除，直到所有的因子被检查完。我们使用 F-值来衡量特征因子的贡献。图 1 描述了加入不同特征因子后关键词抽取的 F-值。初始因子为未加权的 TFIDF。

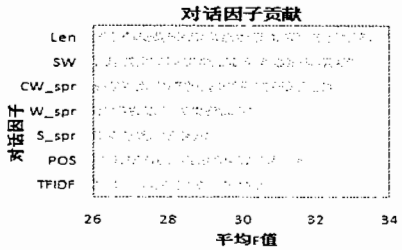

图 1　不同对话因子组合下，关键词抽取的平均 F 值

从图 1 中可以看出加入 POS 信息后稍微改善了 TFIDF 的性能；对于对话者因素，基于句子数目的对话者权重和基于词数的对话者权重都没有带来性能的提升，而只有基于实词数目的对话者权重带来了显著的性能提升。这说明句子数目并不能反映对话者的重要性，这可能是由对话特点造成的，因为对话是自发的、无准备的，所以，可能某个对话者虽然说了很多句子，但是很多句子都非常短，因此，实际说的词数并不多，包含的信息量也就较少，这种现象在对话中是非常普遍的；而基于词数的对话者权重比基于句子数目的权重表现稍好，也说明了这个问题，也即词数比句子数目更有效反应说话者传递的信息量。但是基于词数的对话者权重也没有提高性能，考虑到对话用语比较口语化，有的人虽然说的词很多，但是可能包含了很多虚词、语气词等，因此，词数目也不能很好反映说话者的贡献，而实词可能更能有效地进行反映。基于实词的对话者权重的结果也表明了这点，改种权重提高了系统系能，这表明对话者因素是有益的，同时也说明了实词在信息表达中的重要性；加入句子重要性后系统系能有所改善，支持了摘要句和关键词之间可以相互加强的假设；句子长度因素也提高了系统性能，这表明关键词更可能出现在较长的句子中。

在后面的实验中，我们均采用综合 TFIDF、POS 过滤、基于实词的对话者权重、句子重要性和句子长度这五种因素的加权 TFIDF 进行关键词抽取。

在抽取关键词的基础上，我们进行了二元词拓展。表 2 列出了二元词拓展后系统推荐的 top

标签和人工标注标签的对比示例。其中，黑体字标签是与人工标注完全匹配的标签，斜体字标签是从二元扩展中抽取的二元关键词。整体上，每段对话至少有一个标签推荐正确，而且第一个标签总是与人工标注匹配。同时二元模型也发挥了预期的作用，如表 2 中的系统推荐标签“龙套岁月”和“调查报告”分别是由二元 POS 模板中的“n+n”和“vn+n”扩展得到的。

表2　5个带有标注的对话例子

文档编号	人工标注答案	系统推荐标签
1	《鲁豫有约》, 上山下乡, 龙套岁月, 王学圻, 李雪健, 濮存昕, 空政话剧团	**岁月**, **濮存昕**, **王学圻**, 李雪健, 台词, 宣传队, 话剧团, ***龙套岁月***, 表演
2	生儿育女, 中国, 家庭, 江森海, 中国, 英国, 北京, 胡同串子	胡同, 邻居, **江森海**, 缘分, 女儿, 再生
3	企业, 效益, 质量, 样品, 经济, 调查报告, 玩具生产, 细致, 玩具厂	玩具, 油漆, 分析, 制造, 全球, ***调查报告***, 跌落, ***经济频道***, 注重, 测试, 生产, 油墨, 厂方
4	点击量, Web2.0, 网站, 页面浏览, 流量, 宽带, 风险投资, 赢利模式, 新浪, 编辑	**网络**, **编辑**, 风险, 成本, 客户, 模式, 带宽, 损耗, 商业, 搜狐
5	并购, 欧美业务, 重组, 国际化, 跨国并购, 经验, 时间	**并购**, 修修补补, 决心, 代价, 解决问题, 业务, 经验, 李东生, 指望

注：黑体字表示系统标签与人工标注答案严格匹配；斜体字表示系统标签是一个二元关键词。

为了比较完整的推荐算法的性能，我们使用基于资源协同过滤的 KNN 模型和基于图的 TextRank 模型（Mihalcea and Tarau, 2004；Wu et.al., 2010）作为基线模型进行比较实验。KNN 模型计算文档 Q 和训练文档 Di 的余弦相似度,

$$\mathrm{Sim}(Q,D_i)=\frac{\sum_i n(Q,j)n(i,j)}{\sqrt{\sum_j n(Q,j)^2}\sqrt{\sum_i n(i,j)^2}} \tag{7}$$

其中，n(i , j)代表词 j 在段落 i 中出现的次数。最终，s 个最相似文档中的 top-n 的标签被推荐。

为了评价系统性能，本文使用两种评价策略。第一个是自动评价，使用如下的四个标准对系统的输出与人工标注答案进行对比。

- Top-k 准确度：k 个推荐标签中至少有一个正确的文档占全体测试文档的比例
- Exact-k 准确度：第 k 个标签刚好正确的文档占全体测试文档的比例
- 召回率：推荐正确的标签在人工标签中的比例
- 精确度：推荐正确的标签在所有推荐标签中的比例

本文的标签匹配是基于词的，如果推荐的二元词与标准的人工标注答案没有匹配，那将它切分成单个词与标注答案进行二次匹配。图 2 显示了不同方法在 4,201 个对话段落上的 Top-k 准确度、Exact-k 准确度、精确度和召回率。

从图 2(a)中可以看出，加权 TFIDF 在 top-1 时的准确度达到了 45.89%，加入二元扩展后的 KeyEx 达到了 47.12%，TextRank 和 KNN 的准确度分别为 38.35%和 13.05%；随着标签个数的增加，top-k 准确度的性能逐步得到改善；对于 top-9，KeyEx 可以达到 86.38%的准确度。图 2(b)显示了不同 k 下的 exact-k 的准确度。很明显，KeyEx 在 exact-1 时取得了最好的准确度（47.12%），且随着 k 的增加性能下降。图 2(c)显示了精确度-召回率，KeyEx 具有最佳的性能。随着标签个数从 1 增加到 9，精确度从 42.7%降到 26.27%，而召回率从 19%升到 47.18%。

我们采用的第二个评价方法是人工评测，主要依据拒绝率，它代表有多少推荐标签是不可以被人接受的。通过分析人工标注的不一致性，我们需要质疑，Top-k 准确度、Exact-k 准确度和精确度-召回率是否适合用来评价标签推荐系统的性能？所以我们在少量语料上，采用拒绝率进行人

工评测。我们选取 100 个对话，给 3 个人提供系统输出标签，并让他们标记出哪些标签是完全不能反映对话内容的，然后测量出每个系统/标注者的百分比拒绝率。结果显示在表 3 中。

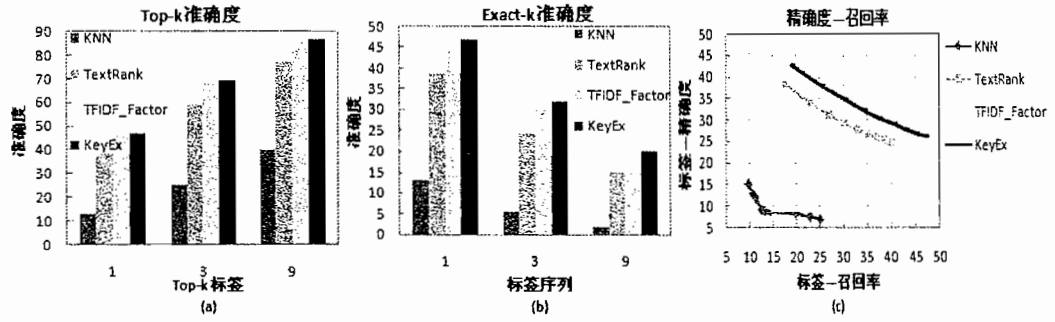

图 2　不同方法在 4,201 段测试对话数据上的性能。(a)Top-k 准确度; (b) Exact-k 准确度; (c)精确度-召回率

表 3　人工评测结果：不同系统/标注者的拒绝率

系统/标注者	拒绝率(%)
Annotator2	6
Annotator1	12
KeyEx	43
TFIDF_Factor	47
TextRank	52
KNN	71

从表 3 中，可以看出，人工标注的拒绝率最低，这不足为怪；总体上，人工评测结果和自动评测结果是吻合的；同时证明了我们的方法对于对话领域的标签推荐是有效的。

4　结论和未来工作

在本论文中，我们提出了一种面向中文对话的标签推荐方法——KeyEx。该方法首先利用 TFIDF 加权同时考虑多种对话因子来抽取关键词，然后利用预先定义的 POS 二元模式进行二元关键词的过滤，最后根据权重推荐出 top-n 的标签。同时，我们利用自动评价和人工评价两种方式对系统性能进行检验。结果显示 KeyEx 优于 KNN 和 TextRank，证明了该方法对于中文对话的有效性。

虽然从评测数据上可以看出，KeyEx 获取的性能远远超过了在网页标签推荐上广泛使用的协同过滤方法（KNN），但是 top-9 的准确度最终也下降到了 26.27%，这与实际应用的需求还相距甚远。通过错误分析，存在的不足之处一方面可能由于语料中存在大量的未登录词，在分词模块中对这些未登录词的识别可能影响到标签抽取的性能，比如："网易"在分词过程中被错分成名词"网／n"和动语素"易／vg"，直接影响到候选关键词提取过程的精度。另一方面，对于关键词、二元词的排序方法过于简单，仅仅考虑词频，未考虑它们的语义关联。因此，在未来的工作中，我们将集中在这些方面的改进和优化。

参 考 文 献

[1] Leandro Balby Marinho, Lars Schmidt-Thieme. 2009. Collaborative Tag Recommendations. Http: //www. springerlink.com/index/m5688 r6761448612.pdf.

[2] Bruno Oliveira, Pável Calado, and H. Sofia Pinto. 2009. Tess: Using resource contents for tag suggestion. In Proceedings of the 5th European Semantic Web Conference.

[3] Ralf Krestel, Peter Fankhauser, Wolfgang Nejdl. 2009. Latent Dirichlet Allocation for Tag Recommendation. RecSys'09, October 23-25, 2009, New York, USA.

[4] S. Brin and L. Page. 1998. The anatomy of a large-scale hyper textual web search engine. Computer Networks and ISDN Systems, pages 107-117.

[5] Radev, S. Blair-Goldensohn, and Z. Zhang. 2001. Experiments in single and multi-document summarization using MEAD. In Proceedings of The First Document Understanding Conference.

[6] Hua-Ping ZHANG, Hong-Kui Yu, De-Yi Xiong, Qun LIU. 2003. HHMM-based Chinese Lexical Analyzer ICTCLAS. 2nd SiGHAN workshop affiliated with 41th ACL; Sapporo Japan, July, 2003, pp.184-187.

[7] R. Agrawal, T. Imieli´nski, and A. Swami. Mining. 1993. Association Rules Betwe'en Sets of Items in Large Databases. SIGMOD Record, 22(2).

[8] Robert Graham, Brian Eoff, James Caverlee. Plurality: A Context - Aware Personalized Tagging System. 2009. Proceeding of the 17th international conference on World Wide Web.

[9] Robert Graham, Robin Burke. Personalized Recommendation in Social Tagging Systems Using Hierarchical Clustering. 2009. Proceedings of the 2008 ACM conference on Recommender system.

[10] Eibe Frank, G.W. Paynter, I.H.Witten, C. Gutwin, and C.G.Nevill-Manning. 1999. Domain-specific keyphrase extraction. In Proceedings of IJCAI, pages 688-673.

[11] Hulth. 2003. Improved automatic keyword extraction given more linguistic knowledge. In Proceedings of EMNLP, pages 216-223.

[12] Kerner, Z. Gross, and A. Masa. 2005. Automatic extraction and learning of keyphrases from scientific articles. In Computational Linguistics and Intelligent Text Processing, pages 657-669.

[13] R. Mihalcea, P. Tarau. 2004. Textrank: Bringing order into texts. In Proceedings of EMNLP.

[14] Wei Wu, Bin Zhang, Mari Ostendorf. 2010. Automatic Generation of Personalized Annotation Tags for Twitter Users. The 2010 Annual Conference of the North American Chapter of the ACL, pages 689-692.