计算机研究与发展
Journal of Computer Research and Development
ISSN 1000-1239/CN 11-1777/TP
49(11): 2344-2351, 2012

Tag-TextRank:一种基于 Tag 的网页关键词抽取方法

李 鹏 王 斌 石志伟 崔雅超 李恒训
(中国科学院计算技术研究所 北京 100190)
(lipeng01@ict.ac.cn)

Tag-TextRank: A Webpage Keyword Extraction Method Based on Tags

Li Peng, Wang Bin, Shi Zhiwei, Cui Yachao, and Li Hengxun
(*Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100190)

Abstract Keyword extraction is to extract representative keywords from texts and has been widely used in most text processing applications. In this paper, we explore the use of tags for improving the performance of webpage keyword extraction task. Specifically, we first analyze the characteristics of bookmarking behavior and find that people usually use the same tags to label multiple topic-related webpages, which is shown by the fact that over 90% of labeled webpages can find relevant webpages through their tag information. Based on the discovery, we propose a method called Tag-TextRank. As an extension of the classic keyword extraction method TextRank, Tag-TextRank calculates the term importance based on a weighted term graph and the edge weight for a term pair is estimated by the statistics of the relevant documents which are introduced by a certain tag of the target webpage. The final importance score for a term is the combination of the above tag dependent importance scores. Tag-TextRank can measure the term relations by utilizing more documents so as to better estimate the term importance. Experimental results on a publicly available corpus show that Tag-TextRank outperforms TextRank on various metrics.

Key words social annotation; tag; keyword extraction; webpage keyword extraction; TextRank

摘 要 关键词抽取是从文本中抽取代表性关键词的过程,在文本处理领域中具有重要的应用价值.利用一种近年来受到广泛关注的新的信息源——社会化标签(tag)——来提高网页关键词抽取的质量.通过对 Tag 数据进行统计分析,发现用户往往对多个在话题上相关的网页使用同样的标签词,一个特定的文档可以通过其标注信息找到相关文档.在此基础上,提出了利用 Tag 进行关键词抽取的框架,并给出了一种具体的实现方法 Tag-TextRank.该方法在 TextRank 基础上,通过目标文档中的每个 Tag 引入相关文档来估计词项图的边权重并计算得到词项的重要度,最后将不同 Tag 下的词项权重计算结果进行融合.在公开语料上的实验表明,Tag-TextRank 在各项评价指标上均优于经典的关键词抽取方法 TextRank,并具有很好的推广性.

关键词 社会化标注;标签;关键词抽取;网页关键词抽取;TextRank

中图法分类号 TP391.3

收稿日期:2010-11-25;**修回日期**:2011-11-15

基金项目:国家自然科学基金项目(60776797,60873166);国家"九七三"重点基础研究发展计划基金项目(2007CB311103);国家"八六三"高技术研究发展计划基金项目(2006AA010105)

文章的关键词通常是指那些能够代表文章的主要内容并能区别其他文章的词汇.关键词抽取作为文本处理的一个基本步骤,广泛应用于文本检索、分类、摘要、专有词典构建及互联网广告等领域[1].

关键词抽取根据抽取对象大致可以分为两类:一类是针对规范文本的抽取,比如对会议或者期刊中的论文进行关键词抽取.这些文章使用的语言规范包含的噪音较少,通常可以利用文档本身的语言学特征和统计特征来实现抽取过程[2-4].另一类是针对网页等非规范文本的抽取.面向网页的关键词抽取主要存在两点不同:一方面,网页中包含很多噪音信息,这阻碍了抽取精度的提高;另一方面,网页本身的结构化信息、网页之间的链接关系以及用户对网页标注的标签信息给网页的关键词抽取提供了一些新的可用的特征.其中,Tag 是用户对网页的描述,但是它们又和关键词有着本质的不同[5].这些信息实际上代表了用户对网页内容的某种理解,直观上可以利用这些信息来辅助提高关键词抽取的效果.这正是本文研究的出发点.

具体地,我们对用户标记行为进行了统计分析,结果表明用户倾向于对多个类似的网页标记相同的 Tag,也就是说同一个 Tag 往往会连接多篇文档.于是,对于要抽取的目标网页,可以根据文档间的这种连接关系扩展出一系列相关网页,这些相关网页可以作为一种辅助资源来计算目标网页中关键词的重要度.这样做显然可以避免单篇目标文档估计时带来的稀疏性问题.对于有多个 Tag 的网页,可以综合基于每个 Tag 给出的关键词重要度得分而得到最后的抽取结果.

1　相关工作

根据所用机器学习方法的不同,关键词抽取方法可以分为无监督及有监督方法.无监督的方法一般利用文档词本身的统计信息来实现关键词的抽取,适用范围广,一些方法可以取得不错的效果.而有监督的方法主要通过挖掘能表征关键词的有用特征来提高抽取的效果,但是对于不同的数据集有监督方法往往需要训练不同的参数.

无监督方法中,一项重要的工作是 2004 年提出的 TextRank[6].其基本思想类似于 PageRank,首先根据词之间的共现关系构建一个词项图,认为词与词之间共现代表一种推荐关系,与重要词共现的词也重要,在这个图上经过迭代可以得到词项的重要度排序.实验结果表明,由于上述词项排序结果是基于稳定的收敛值,因此其效果要好于传统的 TF * IDF 的方法.TextRank 方法在诸如文本摘要等任务中得到了广泛应用[7].但是原始 TextRank 的词项图中词之间的连边没有设置权重.这样最后的关键词输出结果倾向于词频高的词项,而忽略某些出现次数少的专有名词,实际上,对要抽取的文档,由于缺乏足够的统计信息,也很难精确估计词项图中两个词之间的相关度.对于有监督方法,代表性的工作有 Turney 在 2000 年提出的 GenEx[8] 以及 Frank 的 KEA 系统[9].之后的工作通过引入更多的语言学特征(如词性)或者改进分类算法(如使用 bagging 等)来提高效果[2-4].另外,国内一些学者也开展了关键词提取的研究[10-11].

近年来,随着互联网广告的应用发展,网页的关键词抽取逐渐成为研究的关注点.微软的 Yih 等学者[1]考虑到词项在网页中的特征,如是否大小写、是否超链接、是否出现在 html 的 meta 字段或者 title 字段等等,通过逐步加入特征及逐步删除特征,来观察不同特征对最后效果的影响.值得注意的是,在他们的工作中,词项在查询日志中是否出现也作为一种特征加入,最后发现,基于查询日志的特征及 IR 特征如 TF 等对关键词抽取效果提升最大.WWW2009 的一篇文章[12]利用外部知识,如 Wikipedia 来提高抽取效果,通过对要抽取文档进行扩展,构建包含大量 Wikipedia 实体的图,使用社区发现的切割算法来对多主题的文档切分,然后进行关键词抽取,但抽取的词仅限于出现在 Wikipedia 中的实体,并且算法的复杂度很高.实际上,从上面的工作中我们可以得到这样的基本结论:基本的 IR 特征可以找到部分关键词,而加入用户信息或者领域知识则可以进一步提高关键词抽取效果.

本文利用 Tag 信息来提高关键词的抽取效果.近年来,Tag 数据作为一种新的资源,其挖掘和利用已经成为信息检索、社区发现等领域的研究热点.之前研究发现绝大部分的 Tag 在描述对应的标记对象上是准确可靠的,并且同锚文本、查询日志等数据相比,Tag 数据提供了更加独特的信息[13-14].在 Tag 数据的应用方面,文献[15]利用 Tag 进行个性化检索,文献[16]利用 Tag 对检索模型进行平滑,文献[17]利用 Tag 进行文档聚类,都取得了不错的效果.我们认为 Tag 数据有两方面的性质可以利用:一是 Tag 的文本性质,一篇文档的所有 Tag 合并到一起可以看作文档在 Tag 空间的表示;二是 Tag 的

关联性质,即同一 Tag 往往会将多篇主题相关的文档关联起来.通过引入相关文档可以避免仅使用原始文档进行估计的稀疏性,从而提高某些任务的效果,这也是我们提出的基于 Tag 进行关键词抽取的出发点.

2 Tag 特点分析

为了验证 Tag 数据的关联性质,我们对用户标注行为进行了统计,主要是分析用户对某个网页标注的 Tag 中是否至少有一个也是该用户对其他网页的标注结果.如果结论为是或者大部分情况下都成立,那么就可以对目标网页进行扩充.具体地,我们使用了一个公开的 Tag 数据集[18],它也是目前学术界公开的供研究使用的最大的 Tag 数据集.该数据集通过爬取 Delicious 网站内容构建,包括〈User, Tag, URL〉三元组的集合.该数据集包括 142 000 000 个三元组、950 000 个不同的用户以及 45 000 000 个互不重复的网页 URL.

2.1 URL 在用户数上的分布

我们首先统计了在不同用户数目下的网页 URL 分布情况.具体地,我们从上面的所有 URL 集合中随机抽取了约 10% 的 URL 进行统计,统计结果如图 1 所示:

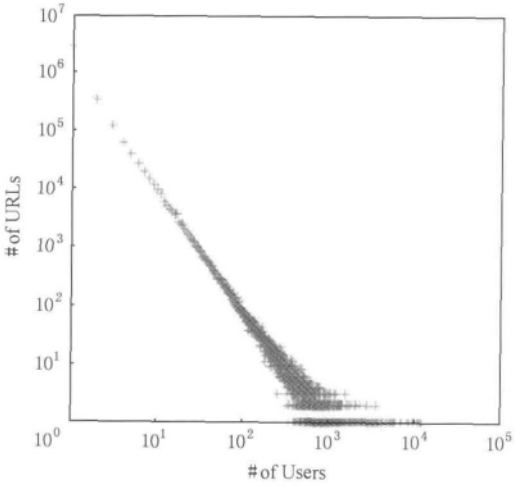

Fig. 1 The distribution of URL frequency.

图 1 URL 在用户数上的分布

图 1 显示出 URL 数与用户数之间呈指数分布(power-law)的关系,随着用户数的增加,URL 数急剧减少.可以看出,只有少数 URL 拥有大量的用户,而绝大部分 URL 只有很少的用户对其标注.具体地,大约有 90% 的 URL,其标注的用户数少于 3.直观上,对网页标注的用户数越多,那么越可能找到相关的网页,对于那些拥有很少用户并且占所有网页很大比例的 URL,是否可以找到相关的 URL,这决定了基于 Tag 进行扩展的方法的适用范围.

2.2 可扩展的 URL 的比例

下面,我们着重分析那些少量用户标注的 URL,它们可以通过 Tag 来扩展相关 URL 的比例,相关 URL 和原始要扩展的 URL 共享同样用户及 Tag.具体地,我们定义:

$\delta(URL)$——对 URL 进行标注的用户集合;

$\pi_u(URL)$——URL 通过对其进行标注的特定用户 $u(u\in\delta(URL))$可以找到的相关 URL 集合;

$\pi(URL)$——URL 的所有扩展 URL 的集合,即通过所有标注用户获得的扩展 URL 的并集 $\sum_{u\in\delta(URL)}\pi_u(URL)$.

我们从 2.1 节中的数据抽取了 5 个 URL 样本集合,每个样本集合有 10 000 个 URL,这些样本集合按照其中文档的用户数进行分类,用户数从 1~5.针对每个样本集合,我们从下列两方面来统计:一是统计可以找到扩展 URL 的原始 URL 的比例,即统计 $|\pi(URL)|\neq 0$ 的 URL 比例;二是统计可以对 URL 进行扩展的用户比例,即 $|\pi_u(URL)|\neq 0$ 的用户的比例.

对于第 1 方面,统计结果如图 2 所示.我们发现随着 URL 用户数的增加,可扩展的 URL 的比例在增大,即使 URL 拥有的用户数为 1,也有超过 90% 的可能性找到相关的 URL.这表明,对于大部分有标注信息的网页,通过用户及 Tag,可以找到与其相关的网页.这验证了 Tag 数据的关联性质,说明基于 Tag 数据扩展相关文档的方法有很好的通用性.

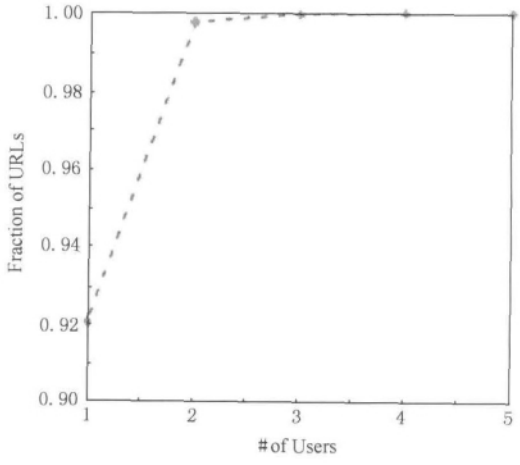

Fig. 2 Fraction of URLs with related URLs.

图 2 可扩展的 URL 比例

对于第 2 方面，统计结果如图 3 所示，我们发现可扩展的用户比例保持稳定，对于不同的样本集合，基本上可扩展的用户占到总体用户的 93%以上，这说明绝大部分的用户倾向于多次使用相同 Tag 来标记文档，也说明了用户的标注行为与网页的流行度关系不大，即与网页被收藏的次数没有关系. 另外，我们进一步观察每个用户扩展的 URL 个数的分布. 由于用户标注行为与网页被收藏次数关系不大，我们统一对 5 个样本集合进行统计，统计结果如图 4 所示. 图 4 的横坐标为扩展的 URL 数，取 $N+1$ 目的是为了画出扩展数为 0 的部分. 显然，扩展的 URL 数的频率服从指数分布(power-law)，0～100 占据大部分.

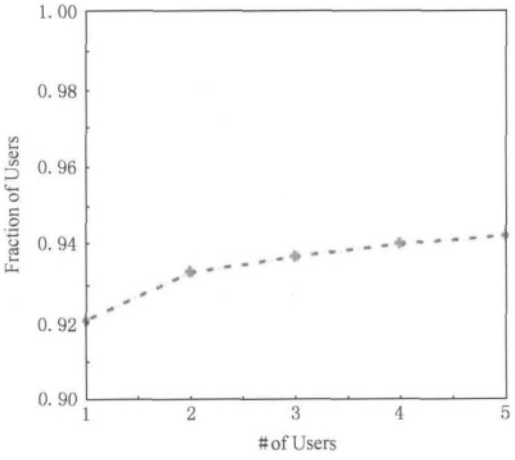

Fig. 3　Fraction of users with related URLs.

图 3　可扩展的用户比例

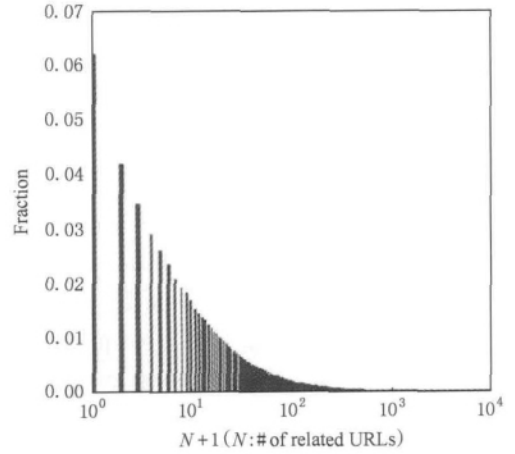

Fig. 4　The distribution of the number of related URLs.

图 4　用户扩展的 URL 个数的分布

3　Tag-TextRank

首先我们给出基于 Tag 的关键词抽取的形式化定义，即给定网页 d，假定其对应的用户标注信息为 $\{\langle user, tag, d\rangle\}$，我们的目标是从 d 中抽取排名最高的 k 个(如 $k=10$)关键词.

为了解决这个问题，我们首先基于词与词在网页 d 中的共现构建一个词项图. 具体地，可以通过设定一个窗口大小，然后将出现在窗口中的任意两个词之间增加一条边. 基于图的排序方法经过迭代可以得到词项的重要度. 在 TextRank 中，不考虑边的权重，认为任意两个词之间的关联度是相同的. 但实际上，在不同主题下词与词之间的关联度可能是不同的. 比如在一篇讲述机器学习的文章中，“machine”和“learning”的关联度会很高，而在讲述如何操作机器的文章中，它们的关系要弱一些. Tag 作为一种高层语义信息反映了文档的主题. 对于一个确定的 Tag，用户标过的相关网页可以作为额外的信息来估计在该主题下，词项间的关联度，这种全局关联度可以作为边的权重来影响最后的结果.

这样对每个 Tag，都可以得到一个词项重要度的排序结果，最后网页关键词的获得可以认为是要综合多个排序结果. 这可以看作是一个数据融合(data fusion)的问题[19].

图 5 给出了基于 Tag 信息进行关键词抽取的框架. 实际上，通过 Tag 引入的相关文档本质上是一种额外的资源，如何利用这些资源进行关键词抽取还可以采用其他方法，比如文献[20]提到的构建更大规模的图. 本文主要探索如何利用该资源对词项的关联度进行计算.

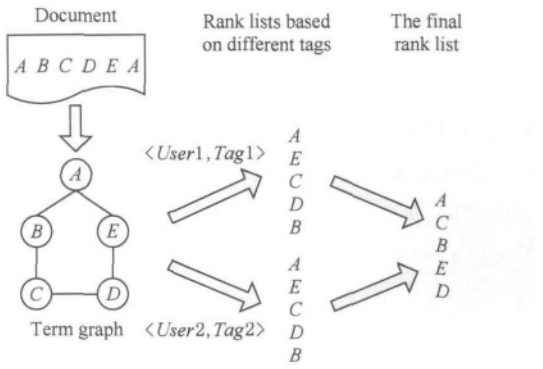

Fig. 5　The keyword extraction process based on tags.

图 5　基于 Tag 的关键词抽取的流程

具体地，我们提出了如下的 Tag-TextRank 算法：

1) 根据要抽取的文档 d 中词的相邻关系，构建词项图 G.

2) 对于 d 上的每一项标注信息 $\langle u, t\rangle$，其中 $u \in U_d$，$t \in T_d$，U_d 是对文档 d 标注的用户集合，T_d 是

文档 d 上的 Tag 集合：

2.1 找到相关文档集合 D^*，相关文档是指与 d 共享相同 User 及 Tag 的文档. 在 D^* 上计算 G 中每条边$\langle w_1, w_2\rangle$的权重 $Meas(w_1, w_2)$，即词项 w_1 与 w_2 的关联度.

2.2 在 G 上运行带权重的 PageRank 算法，获得节点权重的收敛值，输出词项重要度排序结果 $r_{\langle u,t\rangle}$.

3) 合并$\{r_{\langle u,t\rangle} \mid u\in U_d, t\in T_d\}$，生成最终关键词列表 r.

3.1 词项间关联度

词项之间的关联度计算有多种方法可以选择. 基于统计的方法往往在考虑词项单独出现次数的同时考虑它们在某个大小的“窗口”内共现的次数. 常用的关联度指标包括互信息 MI(mutual information)，Pearson's χ^2 统计量及 Dice 系数等等. 前两个指标涉及到窗口总数，而 Dice 系数计算则与窗口总数无关. 这里，我们分别使用 χ^2 及 Dice 系数来计算词项关联度. 计算公式如下：

$$\text{Pearson's} \qquad \chi^2 \frac{(Nn_{ab}-n_a n_b)^2}{Nn_a n_b}; \tag{1}$$

$$\text{Dice coefficient} \qquad \frac{2n_{ab}}{n_a+n_b}. \tag{2}$$

上面公式中，N 表示总窗口数，n_{ab} 为词项 a 及词项 b 共现的窗口数，n_a 为 a 出现的窗口数，n_b 为 b 出现的窗口数. n_{ab}, n_a, n_b 与窗口大小设置有关. 也可以直接从文档层次上统计，即将整篇文档看成一个窗口，这时统计得到的 n_{ab}, n_a, n_b 实际上就是出现的文档数.

3.2 词项排序

通过在 Tag 的相关文档上计算词项间的关联度，最后构建出的词项图为加权无向图. 对于无向图中节点权重的计算，可以将每条边看作是有两个方向的有向图，这样就可以使用带边权重的 PageRank 算法进行求解. 具体的迭代公式为：

$$WS(V_i) = (1-\lambda)+\lambda \times \sum_{V_j\in In(V_i)} \frac{w_{ji}}{\sum_{V_k\in Out(V_j)} w_{jk}} WS(V_j), \tag{3}$$

其中 $WS(V_i)$，$WS(V_j)$ 分别为节点 V_i 和 V_j 的权重，w_{ji} 为节点 V_j 到节点 V_i 的边权重，$In(V_i)$ 为指向 V_i 的节点集合，$Out(V_j)$ 为 V_j 指向的节点集合. PageRank 基于随机游走的模型，到达点 V_i 可以认为有两种途径：一是以 λ 的概率从 V_i 的相邻节点 V_j 跳转；另一种是以$(1-\lambda)$的概率直接到达. λ 的引入也可以看作是为了减少图中的孤立点或者环对节点权重计算的影响.

3.3 结果融合及重排序

结果融合的提出最早是要解决如何合并来自多个检索系统给出的文档序. 类似地，这里要给出词项序的合并结果，考虑到在不同 Tag 下词项得分值之间不具有可比性，所以输入信息主要是词项序信息. 我们使用一种基于序倒数的合并方法[21]：

$$score(w) = \sum_{r_{\langle u,t\rangle}} \frac{1}{rank(w, r_{\langle u,t\rangle})}, \tag{4}$$

其中 $r_{\langle u,t\rangle}$ 是利用$\langle u,t\rangle$的扩展文档得到的词项排序结果，$rank(w, r_{\langle u,t\rangle})$是词项 w 在 $r_{\langle u,t\rangle}$ 中的序. 式(4)考虑了同一词项在多个排序结果中的序，排名一致高的词最后的得分也最高.

3.4 实现及复杂度分析

基于 Tag 的网页关键词抽取使用用户的历史标注信息来获得词项间的关联度，然后利用这种统计关联度来计算词项重要度排序. 在实现上，可以先对 Tag 数据$\{\langle user, tag, d\rangle \mid d\in D, tag\in T, user\in U\}$建立两类索引：一类是从$\langle user, tag\rangle$到文档 d 的索引，该类索引用于快速判断是否存在历史标注信息；另一类是从特定的用户标注的词项对$\langle user, tag, w_1, w_2\rangle$到相应统计信息$(n_a, n_b, n_{ab})$的索引，该类索引用于快速计算词项关联度. 对于新的文档 d 进行关键词抽取，对其每一条标注信息$\langle user, tag, d\rangle$，首先在第 1 类索引中查找是否存在$\langle user, tag\rangle$对应的文档集，如果没有找到，直接使用 TextRank；如果有相应记录，在第 2 类索引中获得词项间的统计信息，这些统计信息结合词项对在当前要抽取的文档 d 中的出现频率，使用指标 $Meas$ 来计算得到关联度. 统计每个窗口的时间是固定的，故建索引的时间复杂度是线性的，为 $O(M+N^*-1)$，其中 M 是历史标注文档集中文档总长度. 在线计算时，假设文档 d 共有 K 条标注信息($\langle user, tag\rangle$)，那么计算的复杂度为 K 倍的 PageRank 迭代的复杂度. 对于大多数网页来说，参照 3.1 节中的分析，90%的情况下，文档的标注用户不会超过 3，并且并不是用户对网页标注的所有 tag 都可以扩展到相关文档，所以 K 一般比较小，相对于 TextRank 不会带来显著的计算量的增加. 实际上，这里的 Tag-TextRank 只给出了一种通用的扩展思路，具体操作中，可以根据需要选择部分标注信息进行扩展，选择的标准可以是针对某个话题的相关 Tag 进行扩展，可以选择高质量的 Tag 等，这可以进一步减少计算开销.

4　实　　验

我们使用了 Bibsonomy[22] 所发布的一个公开语料用于验证 Tag-TextRank 方法的效果. 参照文献[1,12]中的实验方法,我们选择上述语料的博客文章作为最终抽取关键词的目标网页. 具体地,我们随机选取了技术博客网站 zdnet[23] 中的 50 篇博文进行了实验. 在文献[24]中提到,50 个查询上的实验结果具有统计意义,而这里博文充当的就是查询的角色. 在实验过程中,我们对网页进行正文抽取. 经过处理,每篇要抽取的文档平均包含 3.2 个标注信息(⟨*user*,*tag*⟩),在每个标注信息下,扩展后的相关文档数平均为 550 篇. 和之前的统计结果相比,这里用户的标注历史中包含较多的相关文档,对于那些包含较少相关文档的处理依然值得研究.

我们对上述 50 篇文章的关键词进行了人工标注,标注时采用了一种类似于信息检索评价中缓冲池(pooling)的方法[24]. 具体地,我们对要评价的系统(不同参数),各取排在最前面的 L 个词进行求并构成 pool,然后将这些词提供给标注者进行标注. 在我们的实验中,共有 5 个标注者参与了此次标注,最后的标注共产生了 720 个关键词,平均每篇文档关键词数目为 14.4 个,略高于每篇 10 个关键词的最终标注要求,这主要为了避免因为语言差异而导致的关键词遗漏问题. 实验中 L 取 30.

算法涉及到主要参数包括:在抽取文档上构建图 G 所设的窗口大小 N,在 D^* 上统计词项频率所设的窗口 N^*,关联度指标 *Meas*. PageRank 算法中的 λ 使用默认值 0.85.

系统输出的结果和人工标注的结果进行比较,采用的评价指标为正确率、召回率及 F 值. 具体计算公式如下:

$$precision = \frac{|\{\text{人工标注的关键词}\} \cap \{\text{系统的关键词}\}|}{|\text{系统的关键词}|};$$

$$recall = \frac{|\{\text{人工标注的关键词}\} \cap \{\text{系统的关键词}\}|}{|\text{人工标注的关键词}|};$$

$$F = \frac{2 \times precision \times recall}{precision + recall}.$$

4.1　实验结果及分析

我们分别实现了 Tag-TextRank,TextRank 与 TF*IDF 3 种方法. 参数 N 分别取 2,4,6,8,10,而 N^* 取值分别为 2,5,10,20. 词项间关联度计算考虑了 Dice 与 Pearson's χ^2 统计量两种方法. 对于每一种关联度计算,N 与 N^* 的组合会产生 4×5=20 次运行结果.

表 1 列出了 3 种方法的关键词抽取效果. 对 Tag-TextRank 方法,我们列出了在给定参数 N 的情况下,可以达到的最大 F 值及其对应的 N^* 参数.

Table 1　The Comparison of Tag-TextRank, TextRank and TF*IDF

表 1　Tag-TextRank 与 TextRank,TF*IDF 的比较

Method	Parameter			*precision*	*recall*	*F*
	N	*N**	*Meas*			
TF*IDF	—	—	—	0.158	0.133	0.145
TextRank	2	—	—	0.404	0.306	0.348
	4	—	—	0.410	0.311	0.354
	6	—	—	0.411	0.311	0.354
	8	—	—	0.415	0.315	0.358
	10	—	—	0.413	0.314	0.357
Tag-TextRan	2	10	χ^2	0.410	0.300	0.342
	4	20	χ^2	0.417	0.310	0.355
	6	20	χ^2	0.420	0.311	0.357
	8	20	χ^2	0.415	0.307	0.353
	10	20	χ^2	0.408	0.3	0.346
	2	5	Dice	0.412	0.311	0.354
	4	5	Dice	**0.434**	**0.328**	**0.373**
	6	5	Dice	**0.429**	**0.325**	**0.370**
	8	10	Dice	**0.434**	**0.326**	**0.372**
	10	5	Dice	**0.430**	**0.325**	**0.370**

从表 1 可以看出,当词项关联度计算使用 Dice 系数时,Tag-TextRank 的效果要明显好于 TextRank,这说明合理估计词项关联度对最后关键词抽取的作用,同时也表明使用 Tag 的相关文档进行词项关联度估计的合理性. 但是基于 Pearson's χ^2 统计量并没有明显的提高,可能是因为该指标考虑了所有窗口数,而该窗口数会影响词项间关联度的计算.

4.2　参　　数

N^* 是进行词项关联度计算的一个重要参数,N^* 较大,考虑的词项关联范围也越大,一些关联弱的词项可能会被增强,从而带来噪音. N^* 较小,对于图中的边即词项对,它们共现关系很难被捕捉到,从而降低词项关联度估计的准确性. 为此,我们考察在使用 Dice 系数度量词项关联,对于相同 N 值,不同

N^* 对抽取结果的影响. 从表 2 中,可以看出 N^* 一般取 5～10 效果最好.

Table 2 The Sensitivity of Parameter on F-Measure
表 2 不同 N^* 对 F 值的影响

N	N^*				
	2	5	10	20	Doc. Level
2	0.352	0.354	0.351	0.346	0.339
4	0.347	0.373	0.361	0.353	0.348
6	0.346	0.370	0.363	0.360	0.350
8	0.344	0.368	0.372	0.369	0.349
10	0.344	0.370	0.368	0.363	0.362

5 总 结

本文给出了一种利用 Tag 信息对网页进行关键词抽取的框架和方法. 由于 Tag 可以表示用户对文章的理解,反映文章的主题,因此 Tag 信息可以作为一种弱的指导来辅助进行关键词抽取. 我们根据 Tag 数据对目标网页进行扩展,然后在扩展后的网页上估计词项间的关联度. 我们的贡献在于首次利用用户对网页的标注信息进行关键词抽取,最后的输出结果同时考虑了文档表达的多个通过 Tag 体现的主题.

实验结果表明:Tag-TextRank 继承了 TextRank 无监督学习不需要人工标注的优点,在不显著增加计算量的同时,抽取效果要好于后者. 并且前面的统计表明大部分拥有 Tag 的网页可以找到相关网页,这也说明我们的方法有很好的推广性.

用户标注数据的产生及搜集对完善系统有着非常重要的作用,本文提出的方法适用于那些经常被用户使用,并且不断有新的标注数据产生的系统,这可以确保在历史标注信息上计算词项间关联度的准确性. 实际上,本文提出的方法虽然需要用户标注,但是像 Tag 数据的产生首先是用户组织浏览文档的需要,故基于 Tag 的方法不会增加用户额外的操作负担,可以说如何利用挖掘类似于 Tag 的用户数据是未来的信息检索数据挖掘的重要研究方向.

在未来的工作中,我们会尝试评价 Tag 的重要性及用户的权威度,从而选择高质量相关文档来进行关键词抽取.

参 考 文 献

[1] Yih W, Goodman J, Carvalho V R. Finding advertising keywords on Web pages [C] //Proc of WWW'06. New York: ACM, 2006: 213-222

[2] Kelleher D, Luz S. Automatic hypertext keyphrase detection [C] //Proc of IJCAI-05. San Francisco: Morgan Kaufmann, 2005: 1608-1609

[3] Turney P D. Coherent keyphrase extraction via web mining [C] //Proc of IJCAI-03. San Francisco: Morgan Kaufmann, 2003: 434-439

[4] Hulth A. Improved automatic keyword extraction given more linguistic knowledge [C] //Proc of EMNLP'03. Stroudsburg: ACL, 2003: 216-223

[5] Al-Khalifa H S, Davis H C. Folksonomies versus automatic keyword extraction: An empirical study [C] //Proc of IADIS Web Applications and Research 2006. Southampton: ECS, 2006: 132-143

[6] Mihalcea R, Tarau P. TextRank: Bringing order into texts [C] //Proc of EMNLP'04. Stroudsburg: ACL, 2004: 404-411

[7] Wan Xiaojun, Yang Jianwu, Xiao Jianguo. Towards an iterative reinforcement approach for simultaneous document summarization and keyword extraction [C] //Proc of ACL'07. Stroudsburg: ACL, 2007: 552-559

[8] Turney P D. Learning algorithms for keyphrase extraction [J]. Information Retrieval, 2000, 2(4): 303-336

[9] Frank E, Paynter G W, Witten I H, et al. Domain-specific keyphrase extraction [C] //Proc of IJCAI-99. San Francisco: Morgan Kaufmann, 1999: 668-673

[10] Li Sujian, Wang Houfeng, Yu Shiwen, et al. Research on maximum entropy model for keyword indexing [J]. Chinese Journal of Computers, 2004, 27(9): 1192-1197 (in Chinese)
(李素建, 王厚峰, 俞士汶, 等. 关键词自动标引的最大熵模型应用研究[J]. 计算机学报, 2004, 27(9): 1192-1197)

[11] Yang Jie, Ji Duo, Cai Dongfeng, et al. Keyword extraction in multi-document based on joint weight [J]. Journal of Chinese Information Processing, 2008, 22(6): 75-79 (in Chinese)
(杨洁, 季铎, 蔡东风, 等. 基于联合权重的多文档关键词抽取技术[J]. 中文信息学报, 2008, 22(6): 75-79)

[12] Grineva M, Grinev M, Lizorkin D. Extracting key terms from noisy and multi-theme documents [C] //Proc of WWW'09. New York: ACM, 2009: 661-670

[13] Heymann P, Koutrika G, Garcia-Molina H. Can social bookmarking improve Web search? [C] //Proc of WSDM'08. New York: ACM, 2008: 195-206

[14] Bischoff K, Firan C S, Nejdl W, et al. Can all tags be used for search? [C] //Proc of CIKM'08. New York: ACM, 2008: 193-202

[15] Xu Shengliang, Bao Shenghua, Fei Ben, et al. Exploring folksonomy for personalized search [C] //Proc of SIGIR'08. New York: ACM, 2008: 155-162

[16] Zhou Ding, Bian Jiang, Zheng Shuyi, et al. Exploring social annotations for information retrieval [C] //Proc of WWW'08. New York: ACM, 2008: 715-724

[17] Ramage D, Heymann P, Manning C D, et al. Clustering the tagged Web [C] //Proc of WSDM'09. New York: ACM, 2009: 54-63

[18] DAI-Labor. Available data sets[OL]. [2010-06-02]. http://www.dai-abor.de/en/competence-centers/irml/datasets/

[19] Fox E A, Koushik P, Shaw J A, et al. Combining evidence from multiple searches[C]//Proc of TREC-1. Gaithersburg: NIST, 1992: 319-328

[20] Wan Xiaojun, Xiao Jianguo. Single document keyphrase extraction using neighborhood knowledge[C]//Proc of AAAI'08. Menlo Park, CA: AAAI, 2008: 855-860

[21] Zhang Min, Song Ruihua, Lin Chuan, et al. Expansion-based technologies in finding relevant and new information[C]//Proc of TREC-2002. Gaithersburg: NIST, 2002: 586-590

[22] Knowledge & Data Engineering in University of KASSEL. BibSonomy: Dumps for research purposes[OL]. [2010-06-02]. http://www.kde.cs.uni-kassel.de/bibsonomy/dumps

[23] CBS Interactive. Technology news, analysis, comments and product reviews for IT professionals[OL]. [2010-06-02]. http://www.zdnet.com/

[24] Zobel J. How reliable are the results of large-scale information retrieval experiments?[C]//Proc of SIGIR. New York: ACM, 1998: 307-314

Li Peng, born in 1985. PhD candidate. His current research interests include information retrieval and social media (lipeng01@ict.ac.cn).

Wang Bin, born in 1973. Associate professor. Senior member of China Computer Federation. His main research interests include information retrieval: theory and applications(wangbin@ict.ac.cn).

Shi Zhiwei, born in 1973. PhD. His current interests include information retrieval and natural language processing (shizhiwei@ict.ac.cn).

Cui Yachao, born in 1985. Master. Her main research interests include information retrieval and social annotations(cuiyachao@ict.ac.cn).

Li Hengxun, born in 1985. Master. His main research interests include information retrieval and Web crawling(lihengxun@ict.ac.cn).