第36卷 第21期 计 算 机 工 程 2010年11月
Vol.36 No.21 Computer Engineering November 2010
·开发研究与设计技术· 文章编号：1000—3428(2010)21—0260—02 文献标识码：A 中图分类号：TP311

一种基于标签的网页摘要方法

尚书杰，王 灿，朱俊彦
(浙江大学计算机科学学院，杭州 310027)

摘 要：提出一种基于标签的网页摘要方法。根据优质用户和优质标签之间的相互加强关系，利用二分图排序算法对标签进行排序和打分，构建标签-文档图，应用 Manifold Ranking 算法对句子按其重要性进行排序，将排序靠前的句子组成网页摘要。实验结果证明，该方法的摘要准确性有明显改进。

关键词：标签；摘要；图排序

Tag-based Web Page Summarization Approach

SHANG Shu-jie, WANG Can, ZHU Jun-yan
(College of Computer Science, Zhejiang University, Hangzhou 310027, China)

【Abstract】 This paper proposes a two-stage Web page summarization approach by exploiting both the page contents and the tags annotated on that page. Observing the mutually reinforcing relationship between quality tags and quality users, it uses a bipartite ranking algorithm to score tags in the first stage, derives a graph representation for tags and sentences on a Web page and applies the Manifold Ranking algorithm to rank sentences and generates the summary accordingly. Experimental results show that the method has significant improvement in summary accuracy.

【Key words】 tag; summarization; graph-ranking

1 概述

随着网络的迅速发展，万维网成了海量信息的载体。这意味着人们可以从网上获取更多的资源，但同时过多的信息也加重了用户的阅读负担。因此，人们需要一种自动的网页摘要方法，将网页主要内容以简要的方式呈现给用户，以提高用户获取信息的效率。与网页摘要相关的是文本摘要技术。尽管在文本摘要领域已经有较多的研究积累，万维网的发展还是给这一领域带来了新的挑战。与传统的规范化文档相比，网页文档在写作上更加随意，内容组织上更加松散。

传统的文本摘要方法[1]主要针对行文较严谨规范的文档，如新闻、科技论文。在网页文档上直接应用传统的文本摘要方法很难取得理想效果。另一方面，随着 Web 2.0 应用的发展，万维网上出现了大量的用户交互信息，如评论、标签、评分。这些用户交互信息中包含了用户对内容的理解，因此，可以被用作辅助网页摘要的素材。利用用户交互信息辅助网页摘要成了网页摘要领域的一个热门研究话题，相关的工作包括利用用户评论[2]、网页上的标签[3]等来生成网页摘要。但是用户评论容易偏离文章主题，而且评论中可能包含较多噪声。与之相比，大多数用户标签都是用户对相关网页内容的高度概括描述。这使标签成为辅助网页摘要的良好素材。文献[3]对不同类型的用户交互信息与网页主题的相关性进行了调查，调查结果表明，标签比用户评论更适用于网页摘要。本文提出一种基于标签的网页摘要方法，通过挖掘网页上的优质标签以及标签与网页内容的对应关系产生网页摘要。相较于现有基于标签的网页摘要方法[3]，本文方法能够更有效地利用标签中所蕴含的用户兴趣信息，从而使产生的网页摘要更好地聚焦于用户感兴趣的内容。

2 标签评分

专家级用户往往能够更精确地通过标签对相应的网页内容进行概括描述。因此，需要在网页摘要中对来自不同用户的标签进行区分对待。然而，网页上的标签频率没有包含相关的用户信息。同时，网页上的标签呈现长尾分布的特征[3]，即少数标签被大部分用户所标注，而大部分标签只被一两个用户所标注。简单地使用标签频率无法有效地区分大部分标签。为了更有效地使用标签辅助网页摘要，需要对不同用户的标签进行区分。通过观察发现，用户和标签之间存在一种相互加强的关系，即被优质用户标注的标签往往是优质的标签，一个用户给出的优质标签越多，这个用户也更倾向于是一个优质的用户。基于这种观察，本文使用了一种二分图排序算法来有效挖掘优质用户和优质标签之间的这种相互加强关系。图 1 描述了用户与标签之间的标注关系。

图 1 用户-标签图及相应的邻接矩阵

如图 1 所示，左边的结点表示用户，右边的结点表示标签。用户结点与标签结点之间的边表示标注关系，表明一个用户对相应的标签进行了标注。图 1 同时给出了用户-标签图对应的邻接矩阵 $\boldsymbol{M}$。如果用户 i 标注了标签 j，则 $M_{ij}=1$，反之为 0。

本文同时定义用户向量 $\boldsymbol{U}$ 和标签向量 $\boldsymbol{T}$。用户向量 $\boldsymbol{U}=[u_1, u_2,\cdots, u_n]$，$u_i$ 表示用户 i 的对应分值；标签向量 $\boldsymbol{T}=[t_1, t_2,\cdots, t_m]$，$t_i$ 表示标签 i 对应的分值。所有用户的初始分值都设定为 1。在排序过程中，标签得分和用户得分可由式(1)递归地进行计算：

基金项目：国家科技支撑计划基金资助项目(2008BAH26B00)
作者简介：尚书杰(1985－)，男，硕士研究生，主研方向：信息检索；王 灿，讲师；朱俊彦，硕士研究生
收稿日期：2010-04-08 **E-mail：**henochim@zju.edu.cn

$$\begin{cases} \boldsymbol{T} = \boldsymbol{M}^{\mathrm{T}}\boldsymbol{U} \\ \boldsymbol{U} = \boldsymbol{MT} \end{cases} \tag{1}$$

式(1)表明，在每一步迭代过程中，标签的得分等于标注该标签的所有用户得分之和，每个用户的得分等于该用户所标记的所有标签得分之和。式(1)所示的迭代排序过程将最终收敛，相应的标签分值向量 $\boldsymbol{T}$ 将收敛为矩阵 $\boldsymbol{M}^{\mathrm{T}}\boldsymbol{M}$ 的主特征向量。

3 基于标签的网页摘要方法

3.1 标签-文档模型

对文档进行分句是摘要工作中常见的一个处理步骤[1,3]，因此，本文根据标点符号(句号等)将一个网页文档切分为若干个句子，并通过句子间的相似度构建一张句图来表示一个网页文档，接着根据标签与句子之间的对应关系，使用图 2 所示的标签-文档图来更加清晰地描述标签与文档之间的标注关系。

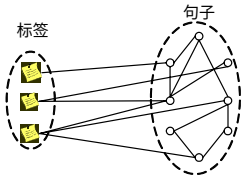

图 2 标签-文档图

在图 2 中，左边的结点表示标签，右边的结点表示句子。标签与句子之间的边表示标签词在句子中的出现情况。若一个标签词在一个句子中出现了 k 次，则该标签与句子之间的边权重被赋值为 k。2 个句子结点之间的边权重对应于其余弦相似度(http://en.wikipedia.org/wiki/Cosine_similarity/)。

用 $\boldsymbol{W}$ 表示图 2 所示的标签-文档图对应的邻接矩阵，则 W_{ij} 可以定义为：

$$W_{ij} = \begin{cases} \text{出现次数, 标签-句子} \\ \text{余弦相似度, 句子-句子} \end{cases} \tag{2}$$

3.2 基于 Manifold Ranking 的摘要生成

本节将标签结点中所蕴含的用户兴趣在标签-文档图中进行传播，并以此对句子结点进行排序，最终选出排序分值高的句子组成网页摘要，从而使产生的摘要有效地聚焦于用户感兴趣的内容。使用 Manifold Ranking 算法[4]来完成上述的排序过程。Manifold Ranking 算法在生成加权图之后，对数据中的若干查询点赋予一个正的分值，其他数据点赋值为零。在排序过程中，所有数据点都通过加权图向它们的近邻传播排序分值，直至达到一个稳态，最终将数据点按它们与查询点之间的相关性进行排序。

Manifold Ranking 算法使用下式对排序分值进行迭代传播：

$$\boldsymbol{F}(t+1) = \alpha\boldsymbol{SF}(t) + (1-\alpha)\boldsymbol{Y} \tag{3}$$

其中，$\boldsymbol{S}=\boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2}$ 是对邻接矩阵 $\boldsymbol{W}$ 的正则化，以保证迭代过程的收敛；$\boldsymbol{D}$ 是一个对角矩阵，D_{ii} 等于 $\boldsymbol{W}$ 中第 i 行元素之和；$\boldsymbol{Y}=[y_1, y_2, \cdots, y_m, \cdots, y_n]^{\mathrm{T}}$ 表示查询向量，设图中共有 m 个标签，则 y_1 至 y_m 对应这 m 个标签的得分，它们的初始值为由第 1 阶段计算所得的标签分值，$\boldsymbol{Y}$ 的其他值均设为 0；$\boldsymbol{F}=[f_1, f_2, \cdots, f_n]^{\mathrm{T}}$ 是排序向量，向量中的每个值 f_i 都对应图 2 中相应结点的排序分值，在初始阶段，$\boldsymbol{F}$ 中的所有值都被赋值为 0。

式(3)中的第 1 部分表示每个结点都把它的分数传播给相邻结点；第 2 部分中的 $\boldsymbol{Y}$ 表示在每次迭代中，每个标签结点都从初始频率中获得一部分分数，从而使每个标签都能最大限度发挥自己的作用。α 是介于 0、1 之间的一个值，用以调节式(4)中两部分的权重。经过不断的迭代后，排序向量 $\boldsymbol{F}$ 的分值将趋于一个稳态，并收敛于下式：

$$\boldsymbol{F}^* = (1-\alpha)(1-\alpha\boldsymbol{S})^{-1}\boldsymbol{Y} \tag{4}$$

式(4)中$(1-\alpha)$是个正的常数，它不会影响排序，因此，最终的排序关系可以由下式表示：

$$\boldsymbol{F}^* = (1-\alpha\boldsymbol{S})^{-1}\boldsymbol{Y} \tag{5}$$

网页文档的摘要由得分最高的若干个句子组成。

4 实验对比

4.1 数据集和评价指标

本文从 MSDN 的 IEBlog(http://blogs.msdn.com/ie/)网站中下载了数百篇文章，并从 Delicious(http://www.delicious. com)网站中下载了这些文章相对应的标签。从这些文章中随机选取了 101 篇作为实验数据集并对文章内容进行了预处理，去除了文章中的停止词，并且根据 Porter(http://www. tartarus.org/martin/ProterStemmer/)的词干提取程序对文章中的单词进行了词干处理，最后根据标点符号(句号等)对文章进行分句。在评价机器摘要的准确性时，机器生成的摘要通常与人工生成的摘要进行对比，人工生成 101 篇文章对应的摘要。在不参考文章对应标签的情况下，从每篇文章中选取 1/3 最重要的句子构成该文章的摘要。本文使用 ROUGE[5]衡量算法生成的摘要质量。ROUGE 是一个被广泛应用的评价摘要准确性的指标。它通过比较计算机生成摘要和人工生成摘要之间的相互覆盖情况(共有词的个数及出现顺序等)计算出机器算法的准确性。本文使用了 ROUGE-1.5.5 程序包，以 ROUGE-1 作为评价指标。

4.2 实验结果

实验对使用以下方法自动生成的网页摘要进行了对比：

(1)OTS(Open Text Summarizer)(http://libots.sourceforge.net/)：一个广泛使用的开源摘要方法，许多新的摘要方法将它作为基准算法进行比较，它仅使用文档内容产生摘要。(2)标签(出现次数)：直接将标签在网页中的标记频率作为得分应用于 Manifold Ranking 公式。α=0.9 时，该方法取得的效果最好。(3)二阶段方法：先由第 1 阶段计算出标签的得分，然后将它们应用于第 2 阶段。α=0.9 时，该方法取得的效果最好。

由表 1 可知，使用标签作为摘要辅助材料的方法比不使用标签而直接对原文进行分析的 OTS 准确度有了明显的提高，其中，对于 F-指标，方法(2)提高了 9 个百分点，方法(3)提高了 12.7 个百分点，这说明标签的确是一个很好的辅助做摘要的资源。在这 3 种方法中，方法(3)的准确度最高，比方法(2)高出 3.6 个百分点。

表 1 3 种方法准确度比较

指标	OTS	标签出现次数	二阶段方法
Recall	0.624 7	0.701 7	0.745 8
Precision	0.623 2	0.708 9	0.739 7
F-指标	0.611 8	0.702 1	0.738 5

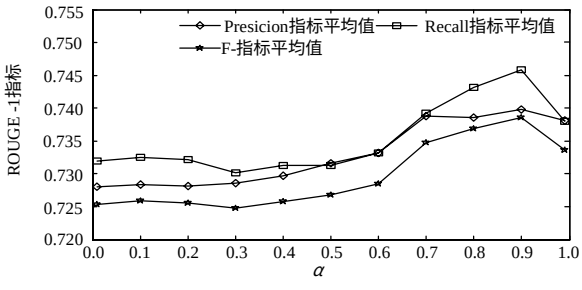

图 3 不同 α 下的 ROUGE-1 值

式(4)中的 α 衡量标签初始频率和相邻结点对最终得分的贡献比重。图 3 为不同 α 值下，ROUGE-1 的 Recall、Precision

（下转第 264 页）