基于关键词抽取的 hypertext 自动建立方法*)

路绪清　唐　杰　李涓子　蔡月茹
(清华大学计算机系　北京100084)

摘　要　随着 Internet 的发展，电子文档的数量成指数级增长，大量的文档之间存在密切的联系。将这些电子文档发布到 WWW 上需要有效地建立这些大量文档之间的链接，从而为用户提供一个更加友好的导航界面。对于以超文本形式产生出来的大量文档，用手工的方式为其指定超链接，不但需要领域知识，而且将是一项极为繁重的劳动。因此，实现超文本建立的自动化是一项很有意义的工作。目前的各种超链建立方法存在着自动化程度不高和准确率低的缺点。本文基于关键词自动抽取提出了一种为文档自动建立超链接的方法。实验证明该方法取得了较好的效果。

关键词　Hypertext，关键词抽取，贝叶斯决策理论，自动建立

Keyword Extraction Based Automatic Construction of Hypertext

LU Xu-Qing　TANG Jie　LI Juan-Zi　CAI Yue-Ru
(Department of Computer, Tsinghua University, Beijing 100084)

Abstract　With the development of internet, electronic document is growing at a skyscraping rate. There is close relation between the large numbers of documents. To publish them to WWW is the requirement of the technical development. It is essential to set up the links between them so that to provide the users a more friendly navigate interface. In this case, large numbers of documents are required to be converted into the form of hypertext. Current approaches to generate hyper links face to the problems of low efficiency and low precision. This paper presents a new approach based on automatic keyword extraction to generate the links of documents automatically. Experiment shows that the result is preferable.

Keywords　Hypertext, Keyword extraction, Bayes decision theory, Automatic construction

1　引言

当前，大量的文档以超文本的形式产生出来，如果以手工的方式为其指定超链接，不但需要领域知识，而且还是一项十分繁重的劳动，这是因为要为文档指定超链，需要首先要理解文档的内容，然后概括出文档的关键词，最后，以关键词为锚点，以另外一篇和这个词有密切关系的文档为目标建立超链。当文档数量急剧增加时，这将是一个极为耗时耗力的过程。

目前有许多方法被用来为文档自动建立超链。文[1]提出了一种综合语义分析和统计方法的超链自动建立方法，在关键词的确定上实现了自动化，但是它采用的方法是一种比较传统的方法，即这种关键词抽取方法依赖于一个预定义的词汇表，当抽取的关键词超出这个词表的范围之后基于这种方法建立的超链将变得不合理。文[2]提出了一种概念框架模型来为文档生成超链，但是这种概念框架模型并没有实现真正意义上的超链自动建立，因为它在指定关键词这一步仍然是用手工指定的。

总的来说，当前为文档建立超链的方法存在两个方面的问题：手工建立费力费时；自动建立方法存在自动化程度不高和准确度较低的弊端。针对这两个问题，本文提出了一种新的超文本自动建立方法 KBACOH(Keyword Extraction Based Automatic Construction of Hyperlink)。KBACOH 实现了超链的完全自动建立，同时在建立过程中兼顾了链接关键词的统计和语义信息[2]。

本文第2节给出了相关研究工作的介绍，第3节详细描述了基于 Bayes 决策的关键词抽取方法，第4节给出了链接的生成算法，第5节进行了实验和相关分析，最后总结全文，并给出了进一步的研究工作。

2　相关工作

当前已经有一些针对超文本自动建立的研究。文[1]给出了一个超链自动建立系统。这个系统建立超链接的过程主要包括关键词抽取，节点生成以及链接的生成等步骤，其基本思想是综合统计和语义相似性来建立超链接。统计相似性方面它使用了 TF * IDF 的加权模式以及向量内积的计算方法来进行关键词抽取；语义相似性方面，使用一种基于预定义词汇表的方法来确认关键词。最后，把两种计算结果加权相加，作为最终的结果。虽然它实现了超链的自动建立，但是它的最大弊端在于在关键词抽取阶段使用了一种比较传统的计算语义相似性的方法，依赖于一个预定义的词汇表，使得这种方法的最严重的缺陷是当抽取的关键词超出词库范围后，很难抽取出正确的关键词。文[2]在关键词抽取阶段考虑到了这一点没有使用词库，它的解决办法是使用手工指定，因此使这个系统没有实现真正的超链建立自动化。

但实际上我们可以使用机器学习的方法来克服传统的关键词抽取过程中受限于预定义词库的弊端，同时还可以实现

*)本文研究得到国家自然科学基金资助(课题编号：60443002)。路绪清　硕士生，主要研究领域为信息检索和半结构化数据处理。唐　杰　博士生，主要研究领域为信息检索与语义 Web。李涓子　副教授，主要研究领域为信息检索与自然语言处理。蔡月茹　副教授，主要研究领域为半结构化数据处理。

抽取的自动化。有关关键词抽取的研究包括：

文[3]给出了一个关键词抽取系统 GenEx，这个系统基于一套确定参数的启发式规则，这些参数通过一个基因算法来调整。基因算法通过这些规则的参数优化了在训练文档集中被正确指定的关键词的数量。Turney 比较了 GenEx 和标准的机器学习技术如 bagged decision trees，GenEx 具有更好的性能。同时还说明了 GenEx 的一个重要特征，就是它通过一个杂志论文的文档集进行训练能成功地从不同主题的网页上抽取关键词。

和 GenEx 相比，另一个关键词抽取系统 KEA[4]提高了一些性能，特别是当它的训练集和测试集都来自同一个领域的时候。由于没有使用遗传算法，KEA 加速了训练和抽取的过程。在 KEA 中，对关键词特征的描述使用的是 TF * IDF 和短语在文中首次出现的位置。

本文提出了一种基于贝叶斯决策理论的关键词抽取方法，我们选取的特征较之 KEA 更兼顾了词的语义方面的特征。

3 关键词抽取

基于贝叶斯决策理论的关键词抽取方法，能够有效地实现关键词的自动抽取，可以进一步实现超链的自动建立。下面分别介绍这个抽取系统的特征选择和抽取模型。

3.1 特征的选取

在机器学习算法中，特征选取是极其重要的一步，好的特征能够更加正确地代表单词本身，还能导致更好的映射。我们的方法中选取了四个特征：词和文档之间的互信息，词的上下文，词之间的关联关系，在文档中的首次出现的位置。这四个特征很好地兼顾了词在文档中的统计和语义信息，较之文[3]中独立的计算统计相似性和语义相似性然后再加权相加的方式，本方法具有一定的优越性。

代表单词的特征包括四个属性：互信息，文档中的首次出现的位置，词的上下文，词的连接。下面给出每个特征的计算方法。

1. 互信息　互信息在信息检索中是一个标准的指标[6]，它表示单词和文档互相提供的信息量。定义如下：

$$w_{i,j}=\frac{tf_{i,j}/NN}{\frac{\sum_i tf_{i,j}}{NN}\cdot\frac{\sum_j tf_{i,j}}{NN}}\times factor$$

$$factor=\frac{tf_{i,j}}{tf_{i,j}+1}\times\frac{\min(\sum_i tf_{i,j},\sum_j tf_{i,j})}{\min(\sum_i tf_{i,j},\sum_j tf_{i,j})+1}$$

其中：$tf_{i,j}$是词 i 在文档 j 中出现的次数；$NN=\sum_j\sum_i tf_{i,j}$是所有单词在所有上下文中总的频数；$factor$ 是用来平衡互信息量偏向非频繁词的误差。

2. 首次出现位置　即被计算为在候选词第一次出现前的词的个数，用它除以该文档词的总数。这个值是一个界于0和1之间的实数。

3. 词的链接　也是一个重要的特征。这里给出两个连接的定义：linkage authority 和 linkage hub。前者表示一个词在多大程度上被其他词改变，改变这个词的词越多，这个词的权值越高；后者表示多少个词改变了其他词，改变词越多这个词就具有越高的 hub。因此，定义连接如下：

$$wl_h=\frac{freq(w_i,\forall)}{count(\forall,\forall)}\times-\log\frac{df(w_i,\forall)}{N}$$

$$wl_a=\frac{freq(\forall,w_i)}{count(\forall,\forall)}\times-\log\frac{df(\forall,w_i)}{N}$$

其中：wl_h 和 wl_a 分别代表修改关系和被修改关系；$freq(w_i,\forall)$ 是 w_i 在给定的文档中所修改的词的数量；$df(w_i,\forall)$是在全局集合中包含修改关系；$freq(w_i,\forall)$的文档的数量；$count(\forall,\forall)$是修改关系的总的数量；N 是全局文档集合的数量。

wl_h 和 wl_a 取值在 0 和 1之间。$\log(df(\forall,w_i)/N)$是这个短语出现在文档集合中任何文档中的可能性的对数值。

4. 词的上下文权值　经验分析表明，词的上下文对关键词的抽取也将有很大的作用，词的上下文可以被定义为在这个词前后指定个数的词。实验中，我们指定词的个数为20个单词，包括其前面的十个词和之后十个词。基于这个定义，计算词的上下文的公式如下：

$$wc_i=\sum_j mi_j/size(c)$$

其中：$\sum_j mi_j$ 是上下文中词的交互信息的数量；$size(c)$ 是上下文中词的数量。

实验分析表明，关键词上下文总是包含具有高 TF * IDF 值的词。

3.2 抽取模型[7]

我们的抽取模型基于贝叶斯决策理论，这为解决不确定的行为和推理问题提供了一个坚实理论基础[5]。贝叶斯决策理论描述如下：假设 x 由一个随机变量 X 产生，这个随机变量的分布依赖于参数 θ。令 $A=\{a_1,a_2,\cdots,a_n\}$代表所有可能的行为。进一步假定和每个行为以及每个 θ 值相关联的损失函数 $L(a_i,\theta)$，这些 θ 值即是决策参数。接下来的任务是决定采取哪个行为。

为了评估每个行为，考虑和贝叶斯期望损失相关联的行为 a_i 给出如下具体的 x：

$$R(a_i|x)=\int_\theta L(a_i,\theta)p(\theta|x)\mathrm{d}\theta$$

贝叶斯决策理论指出对于给定 x，最优决策就是选择具有最小期望风险的行为：

$$a^*=\arg_a\min R(a|x)$$

实验中我们所关心的是选择一个最优的词集合作为关键词集合。词是否是关键词依赖于四个特征：交互信息 mi，词的上下文 wc，词的连接 wl，在文档中的首次出现位置 fp。因此行为 a_i 的期望风险为：

$$R(W|DC,D)=\int_\theta L(W,\theta,D,DC)p(\theta|DC,D)\mathrm{d}\theta$$

$\theta\equiv(\{\theta_i\}_{i=1}^N)$，后验分布为 $p(\theta|DC,D)\propto\prod_{i=1}^N p(\theta_i|cw,mi,fp,mr)$

在这种情况下，根据风险最小化理论得到下述公式：

$$W^*=\arg_w\min R(W|DC,D_i)$$

也就是说，选择 W 作为关键词。为了简化计算选择信息损失函数为：

$$L(W,\theta,D,DC)=\sum_{w\in W}-\delta(w,D)$$

如果 w 是一个关键词，$\delta(w,D)=1$；否则 $\delta(w,D)=-1$。

根据以上描述的简单贝叶斯理论，一个词是关键词的可能性取决于它的四个特征，公式如下：

$$p(y|mi,wc,fp,wl)=\frac{p(mi|y)p(wc|y)p(fp|y)p(wl_a|y)p(wl_h|y)p(y)}{p(y)+p(n)}$$

其中：$p(mi|y)$，$p(wc|y)$，$p(fp|y)$，$p(wl_a|y)$，$p(wl_h|y)$是先

验概率，可以根据训练文档来计算。$p(y)=\frac{Y}{Y+N}$是训练文档集中正实例出现的概率，Y 是正例的个数，N 是反例的个数。这样，所有的候选词可以通过这个后验概率排序。同时应用几个剪枝规则对所得的结果进行剪枝。包括：如果两个词有相同的值，则看其词性标注，例如，名词将比别的词优先。其次，在候选词集中包含词数量越多的可能性越大，例如，如果两个排列位置比较靠前的单词有交叉，其中一个是另一个的一部分，那么这个具有较低 mi 和 wc 值的词将被删掉。最后，前 k 个词被返回，k 是需要抽取的关键词数量。

3.3 关键词抽取流程

关键词提取有训练和提取两个主要阶段：

1. 预处理　首先规范化文档，确定候选词/词组。这个正规化过程包括分词、去除停用词、词干提取。在这个过程中每个词的频率将被计算出来，下一步将用到这个频率，关于分词，采用3-gram 策略。

2. 权值　根据在前一步搜集的信息，为每个候选词计算它的四个特征值。

3. 离散化　所有的特征值都是实数，机器学习之前将其转化成 nominal data 是必要的。训练过程中，基于训练数据，可以生成一个离散表，这个表为每个特征给出了一个数值的范围，每个特征的值将被其落入其中的这个数值范围代替。

4. 先验分布　为了简化提取阶段，为每个属性计算先验分布是必要的。计算 $p(mi|y)$，$p(wc|y)$，$p(fp|y)$，$p(wl_a|y)$，$p(wl_h|y)$和 $p(y)$，$p(n)$。至此，训练阶段完成。

5. 后验分布　在提取阶段，测试集中的每个文档也需要首先进行步骤1～3的操作。然后，对于每个候选关键词，计算其后验分布 $p(y|mi,wc,fp,wl_a,wl_h)$。

6. 最小风险提取关键词集合　基于贝叶斯决策理论，选择具有最小风险的词集合。

4 生成链接

对于文档集合中任意文档都可以表示为以下四元组：

$DC=\{Ds,Keys,Hypers,Constraints\}$

其中：Ds 是所有文档的集合；$Keys$ 是所有关键词的集合，满足$\{\exists\ D_i,key_k\in D_i|key_k\in Keys,D_i\in Ds\}$；$Hypers\in D\times D$ 是文档之间超链的集合；$constraints$ 是超链建立约束的集合。

考虑到作为一定程度上代表这篇文档的关键词一旦出现在另一篇文档中，就在这个词和以这个词为关键词的文档之间建立一个连接，这将会出现许多无意义的连接。因此在连接的过程中使用词与文档的互信息 mi 作为连建立的标准。在建立链接的过程中，将会遇到两种情况。

情形1：如果关键词 $key_{i,j}$ 是文档 D_i 的一个关键词，而这个词还出现在了另外一篇文档 D_m 中，如果 $key_{i,j}$ 在 D_m 中仍然作为关键词，那么立即将其作为 D_m 中的一个锚点，指向 D_i，反之亦然。

情形2：如果 key_k 不是文档 D_i 的一个关键词，那么就要考虑是否将其作为锚点。这种情况下，如果这个词和文档 D_i 的互信息大于一定的阈值 α，即表明这个词和为文档 D_i 提供了一定量的信息，它和文档 D_i 也具有很紧密的关系（虽然它不是 D_i 的关键词），就为其建立到文档 D_i 的链接。在使用互信息作为标准建立链接的过程中阈值的确定我们采用了一种“相对阈值”的策略，策略描述如下：

假设 key_k 为文档 D_i 的一个关键词，同时 key_k 在文档 D_1，D_2，…，D_m 中出现但并不是这些文档的一个关键词，它与这些文档的互信息量为 $mi_1,mi_2\cdots,mi_m$，我们将这个互信息量序列按非递增排序，得到线形序列 $MI_1,MI_2,\cdots,MI_m$，如果在某个值之后的互信息量值骤然下降，则此值之后的值均舍去，仅为此值之前那些值代表的文档建立这个词代表的锚点。例如，对于互信息量序列0.8，0.7，0.6，0.05，0.04，0.001，互信息量值为0.05之后的文档均不为其建立与这个词相关的链接，即当下降幅度大于某个百分比时，就将这个值之后的值均舍去，在实际操作过程中这个值取为：$\alpha=0.7$。

下面是实验中自动建立超链的一个例子：

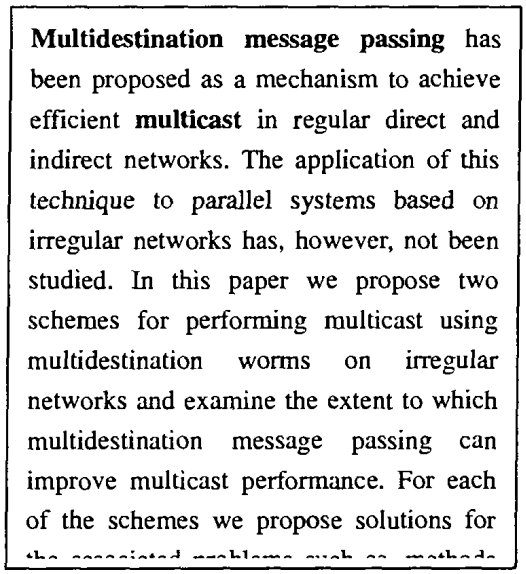
Multidestination message passing has been proposed as a mechanism to achieve efficient **multicast** in regular direct and indirect networks. The application of this technique to parallel systems based on irregular networks has, however, not been studied. In this paper we propose two schemes for performing multicast using multidestination worms on irregular networks and examine the extent to which multidestination message passing can improve multicast performance. For each of the schemes we propose solutions for

图1　关键词所在的文档片断

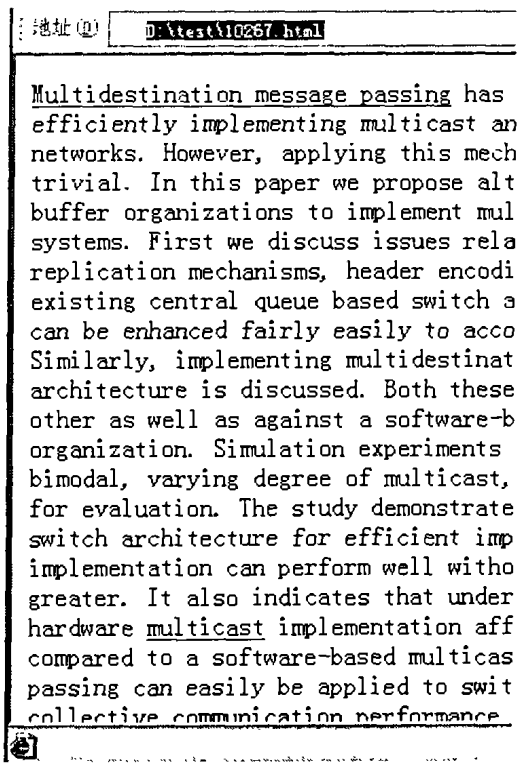

图2　根据图1所示文档中的关键词在本图所示的文档中建立链接

在这个例子中，Multidestination message passing 和 multicast 都是图1所示的文件的关键词，其中 multicast 又是图2

（下转第228页）

将在特征选取和反馈方法及反馈速度上作进一步研究。

参考文献

1 Fish J, Scrivener S. Amplifying the mind's eye: Sketching and visual cognition [J]. Leonardo, 1990, 23(1): 117～126

2 孙正兴，徐晓刚，孙建勇，金翔宇. 支持方案设计的图形输入工具. 计算机辅助设计与图形学学报，2003，15(9)：1145～1152，1159

3 周若鸿，孙正兴，张莉莎，徐晓刚. 草图理解技术研究进展. 计算机科学，2004，31(4)

4 Liu W Y, Jin X Y, Sun Z X. Sketch-Based User Interface for Inputting Graphic Objects on Small Screen Devices, Lecture Notes in Computer Science, Springer, 2002, 2390: 67～80

5 Rocchio J J. Relevance feedback in formation retrieval. In : The Smart Retrieval System: Experiments in Automatic Document Processing, Gerard S, ed, 1971. 313～323

6 孙建勇，金翔宇，孙正兴. 一种快速在线图形识别于归整化方法. 计算机科学，2003，30(2)

7 Xu X G, Sun Z X, Liu W Y, Matching Spatial Relation Graphs Using a Constrained Partial Permutation Strategy. Journal of Southeast University, 2003, 19(3): 236～239

8 Salton G, McGill M J. Introduction to Modern Information Retrieval. McGraw-Hill Companies, March 1984

9 Huang J, Kumar S R, et al. Image indexing using color correlograms. In: IEEE conf. on computer vision on Pattern Recognition, 1997. 762～768

(上接第195页)

中文本的一个关键词，但 Multidestination message passing 不是图2中文本的关键词。Muticast 之所以建立超链是因为它既是第一篇文档的关键词又是第二篇的关键词，而 Multidestination message passing 到文档一的超链是根据互信息量决定的。

如果相似度的值大于阈值 α，则创建一个从这个锚点关键词指向节点的一个链接。因此，有可能出现从一个锚点关键词产生指向数个节点的链接，这种情况将比目前 Web 上的链接形式有更大的便利，因为用户可以自由地选择他想访问的节点，同时这种方式还能使用户高效地访问超文本，因为用户不需要访问那些他不想访问的节点，而这种情况在目前的 Web 中是不可避免的。

5 试验和分析

这里主要给出实验结果和分析。到目前为止，对于关键词抽取还没有一种标准的评估方法。因此，我们通过比较人工指定的关键词和用 KBACOH 指定的关键词来评价基于关键词抽取的 hypertext 建立方法。本文主要采用查准率(precision)来评估实验结果

$$Precision = |kw_m \cap kw_d| / |kw_d|$$

其中：kw_d 是用机器方法抽取的关键词集合；kw_m 是人工指定的关键词的集合。

实验的评估主要集中在分析 KBACOH 自动生成的超文本在多大程度上符合人手工指定的超文本。实验基于两个文档集合共420篇文档。由实验室的5个同学手工指定一定数量文档的关键词用来评估。测试实验结果如下表所示：

表1 训练及测试集合

文档集合	文档总数	测试文档数	训练文档数
CSTR_abstracts	180	120	60
CRANFIELD	240	150	90

表2 手工指定的超链和自动建立的超链的对比

文档集合	手工指定的链接个数	匹配个数	准确度(%)
CSTR_abstracts	5	2.7	54.0
CRANFIELD	6	3.5	58.3
平均	5.5	3	54.6

手工指定的超链由于主观想法和个人理解力的一些影响，也不能保证所有的超链都十分合理，所以，可以说 KBACOH 方法的实际效果应该比我们的试验结果还要好，由此看出 KBACOH 既实现了超链建立的自动化，同时还获得了很高的精度。

结论 大量文档以超文本的形式出现，其手工指定超链接势必成为一项繁冗的体力劳动。因此自动为文档建立超链成为一项十分有意义的工作。在自动建立超链的过程中，关键词抽取是最重要的一步，不同于传统的 IR 过程，在衡量超链的建立情况时，查准率比查全率更重要。本文使用一种基于贝叶斯决策理论的机器学习方法来为文本抽取关键词，进而建立超链接。这种基于学习的方法既实现了自动化，又避免了传统的依赖与词库的关键词抽取方法的缺点。最后在 CSTR_abstracts 和 CRANFIELD 两个文本集上进行了实验。实验结果取得了较好的结果。

在此工作基础上我们将要做的工作是，进一步调整或改变学习过程中的一些特征值参数以便进一步提高抽取关键词的精度，另外还要进一步细化节点的粒度，用本文中的方法在更小的粒度上为文本建立超链。

参考文献

1 Shin D, Nam S, Kim M. Hypertext construction using statistical and semantic similarity. In: Proc. of the second ACM intl. conf. on Digital libraries, July 1997

2 Agosti M, Crestani F. A methodology for the automatic construction of a hypertext for information retrieval. In: Proc. of the 1993 ACM/SIGAPP symposium on Applied computing, March 1993

3 Turney P D. Learning to extract keyphrases from text: [Technical Report ERB-1057]. National Research Council, Institute for Information Technology, 1999

4 Frank E, Paynter G W, Witten I H. Domain-Specific Keyphrase Extraction. In: Proc. of the 6th Intl. Joint Conf. on Artificial Intelligence (IJCAI-99), Stockholm, Sweden, Morgan Kaufmann. 1999. 668～673

5 Berger J. Statistical decision theory and Bayesian analysis. Springer-Verlag, 1985

6 Pantel P, Lin D. Discovering word senses from text. In: Proc. of the eighth ACM SIGKDD intl. conf. on Knowledge discovery and data mining, July 2002

7 Tang Jie, et al. Loss Minimization based Keyword Distillation. APWeb 2004(accepted)