手绘草图识别中的相关反馈方法*)

王 强 孙正兴 李曼舞 蒋 维
(南京大学计算机软件新技术国家重点实验室 南京210093)

摘 要 手绘草图是人类思维外化和表达设计意图的有效工具之一,手绘草图的模糊性和用户适应性问题是草图识别中的关键问题。本文提出了将相关反馈机制引入到手绘草图识别中以捕捉用户意图的方法,该方法以抽取手绘草图的向量化特征为基础,首先利用基于图形特征的相似度计算,给出手绘草图候选识别结果集,然后借助用户对识别结果的相关性评价,通过逐渐调整图形构成特征的权重来捕捉用户输入意图,并提高识别效果。实验表明本文所提出的方法具有很好的效果。

关键词 草图识别,特征向量,相关反馈,权重调整

Relevance Feedback in Freehand Sketchy Shape Recognition

WANG Qiang SUN Zheng-Xing LI Man-Wu JIANG Wei
(State Key Laboratory for Novel Software Technology, Nanjing University, NanJing 210093)

Abstract Freehand sketching is an efficient way for recording and conveying ideas, especially in the early stages of design and idea externalizing. Main challenge in this area comes from the ambiguity of sketch and the user adaptation of sketch recognition. In this paper, we adapt the relevance feedback for sketchy shapes recognition to capturing user's intents. Firstly, the candidate sketchy objects are extracted by means of similarity calculation, based on the establishment of feature-based vector model of freehand sketches. Secondly, the relevance feedback are used to capture the users' input intends and refine the recognition results incrementally by re-weighting the feature-based vector model alone with the user's relevance judgement. Experiments prove the proposed method both effective and efficient.

Keywords Sketch recognition, Feature vector, Relevance feedback, Re-weighting

1 引言

手绘草图是人类一种自然而直接的思路外化和交流方式[1]。手绘草图识别的难点是如何根据人对形状的视觉认知规则、思维习惯和信息模型对用户输入草图及其过程进行猜测和推理,进而捕捉和理解用户意图,来模拟人类"纸-笔"交互模式[2]。

手绘草图的最重要特征是"模糊性",这种模糊性使得手绘草图具有强大的信息表达能力,利于创造性思想的快速表达、抽象思维的外化和自然交流,但同时,这又为手绘草图识别带来很多困难。这个困难最终可归结为输入草图形态与人的视觉感受之间差异的"感觉鸿沟"及输入草图所表达的概念与人的理解之间差异的"语义鸿沟"这两个方面,本质上是手绘草图识别算法如何有效捕捉用户的输入意图问题。尽管国内外对手绘草图进行了广泛的研究[3],但已有的手绘草图识别方法基本上是基于几何的,即采用几何相似性计算方式或基于产生式规则表达,并采取适当的用户选择交互方式来完成草图的识别[2,4]。一方面,几何特征是知识的载体,但它并不能完全表达知识,另一方面,在人类视觉思维规律问题解决之前,这种具有创造特征的模糊性是无法靠仅仅利用算法来消除的。因此,借助于用户的反馈来完成从模糊信息到计算机能够处理的精确信息的映射是手绘草图识别中必须包含的方式之一。除了模糊性特点以外,草图识别过程是与用户自身习惯、特定领域和标准图形库密切相关的,不同的识别之间用户可能会变化,领域、图形库也可能也是变化的,甚至同一用户的习惯都会发生变化,每一次识别都具有自身的内涵特性,这就是草图识别中的用户适应性问题,传统的草图识别方法很难解决用户适应性问题。

相关反馈技术源于文本检索系统,已经有30多年的历史[5]。在文本检索中,相关反馈指的是用户向搜索引擎对检索结果给出相关性反馈信息,搜索引擎通过用户反馈去构造一个更好的查询描述,然后利用新的描述重新计算检索结果。它将检索过程分割为更小的检索序列,便于逐步逼近用户的意图。本文试图以图形特征的向量化为前提将相关反馈技术引入到手绘草图识别过程中,在用户交互的过程中逐步获取用户的意图并挖掘出每次识别的内涵特性,最终识别出用户理想的目标图形。

2 手绘草图识别的相关反馈方案

本文设计了如图1所示的引入相关反馈的手绘草图识别方案。该方案包含两大部分:一是图形识别。在识别出基本图元识别[6]的基础上抽取出图形的向量化特征,包括边类型特

*)本文得到国家自然科学基金项目(编号:69903006、60373065)资助。王 强 硕士生,研究方向为智能用户接口。孙正兴 教授,博导,研究方向为智能用户接口、多媒体挖掘和图形图像技术。李曼舞 硕士生,研究方向为草图信息管理与检索。蒋 维 硕士生,研究方向为智能用户接口。

征向量和简单空间关系特征向量。然后对手绘草图和标准图形库中图形进行相似度计算并给出候选识别结果集。二是用户反馈。首先由用户对候选识别结果集进行相关性评价和标注，然后通过权重调整相关反馈机制对权重进行适当调整，不断提高识别效果最终识别出用户希望得到的图形。有关基本图元识别部分可参见文[6]，本文重点介绍有关特征抽取、相似度计算以及用户反馈的实现过程和方法。

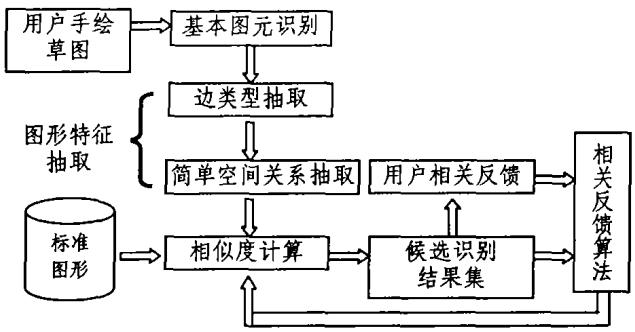


图1　采用相关反馈的手绘草图识别方案

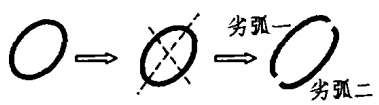
3　图形特征抽取

相关反馈算法是通过调整向量的权重来实现的，在相关反馈之前，首先需要对基本图元进行简化，然后提取图形的构成特征并采用向量模型来表示。

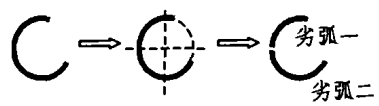
3.1　基本图元简化

基本图元识别过程将用户输入的笔划经过图形预处理、特征识别、图形拟合和规整等过程，自动而即时地识别出基本图元[6]。基本图元包括椭圆、弧和直线3类图元，而弧分为优弧和劣弧，所以实际需要处理4种图元。在4种基本图元基础上抽取的特征向量的维数是较高的(75维)，因此从算法时间和空间效率上的考虑，我们把4种基本图元简化为2种：直线和劣弧，这样可以大大减少特征向量的维数。最重要的是这种简化没有损失草图信息，不影响图形识别的效果。

基本图元简化主要包括：优弧简化为劣弧和椭圆简化为劣弧。对于椭圆，在基本图元识别过程中已经获得它的两个轴，就按照较长轴的方向把椭圆分为两个劣弧，见图2(a)；对于优弧，我们先把它拟合到椭圆，然后按照分割椭圆的方法分割优弧为两个劣弧，见图2(b)。



(a) 椭圆的分割



(b) 优弧的分割

图2　基本图元简化

3.2　特征抽取

(1)边类型特征：每个图元都是图形中的一条边，我们在此基础上根据图元类型和大致方向抽取出图形的边类型特征。在基本图元简化后图形中只存在直线和劣弧两种图元。我们按照直线的四个大致方向把直线分为4种类型，如图3(a)；对于劣弧，按照劣弧对应弦的大致方向和弧的朝向把劣弧分为8种类型，如图3(b)。因此共存在12种边类型。

(a) 四种直线　　(b) 八种劣弧

图3　12种边类型

(2)简单空间关系特征：关系可以用(a,b)表示，其中a,b为关系中的两条边。图形中所有的关系(a,b)组成了图形的简单空间关系。这里面的边就是指直线和劣弧，因此存在直线和直线、劣弧和劣弧、直线和劣弧三种组合方式。简单空间关系主要从关系角度考察图形，而对组成关系的两条边的类型不敏感。两条边之间关系根据交点个数和总体形状一共被分成12种类型，如图4。

图4　12种关系类型，单元格中为相应类型的举例

特征抽取过程中，边类型特征具有平移和缩放不变性，关系类型特征具有平移、缩放和旋转不变性。经过边类型和简单空间关系抽取后，需要对特征向量化，我们用一个12维的 *Edgetype* 向量表示图形的边类型特征，其中 *Edgetype* [*i*]存放图形中边类型为 *i* 的边的个数。用一个12维的 *Relation* 向量表示图形的简单空间关系特征，其中 *Relation* [*i*]存放图形中关系类型为 *i* 的关系个数。

3.3　复杂度分析

假设 *n* 是草图中边的个数，对于该特征抽取所需空间主要是存放图形中的边，至于特征向量是两个定长的12维向量，可以忽略不计，所以算法的空间代价为 $O(n)$。对于抽取边类型特征需要遍历 *n* 个边；对于抽取关系类型需要找出 *n* 条边中任意两个边的关系，需要访问 C_n^2 次边，所以时间复杂度为 $O(C_n^2+n)$。综上所述，算法空间复杂度为 $O(n)$，时间复杂度为 $O(n^2)$。

4　相似度比较

特征抽取后，图形就有了特征向量 *Edgetype* 和 *Relation*。因为标准图形库中图形的特征向量是预先抽取的，此时就可以对用户输入图形和标准图形库中图形进行基于特征向量的相似度比较，过程主要包含下述7个步骤，其中第5,6,7步

的具体方法在第5部分描述。

Step 1 初始化权重：$W=[W_i, w_{i\_j}](i=1,2, j=1,2,\cdots,12)$，其中 W_1, W_2 分别为边类型特征和简单空间关系特征的权重；w_{i_j} 为特征向量中每个维度上的权重。至于初始权重的设置，对于 W_1 和 W_2 都存在一定的经验值。实验中，我们发现关系特征相对边类型特征对于图形比较重要，因此我们赋予 W_2 较高的初始值，我们取 $W_1=1/3$、$W_2=2/3$。对于两个特征向量中每个维度上的权重 w_{i_j}，我们就简单地把初值均设为1/12。

Step 2 分别计算每个特征的相似度 Sim_E 和 Sim_R。在计算相似度之前，需要对向量做归一化操作。归一化的目的是让特征向量的各分量在相同的范围内取值（特征内部归一化），以及让每个特征相似度在相同的范围内取值（特征间归一化）。

对于特征内部的归一化，在边类型特征中所有分量的取值范围在理论上是一样的，因此无须作归一化。在简单空间关系特征中，平行关系具有传递性，因此 *Relation* [2]的平均值比其它关系类型的平均值要大，为此需要平滑平行关系的值，我们用一个对数变换来平滑这个值：$Relation\ [2]=LOG^{Relation[2]}$。

相似度距离采用经典的二次欧拉距离。图形 A 和 B 的两个特征的相似度距离分别为：边类型特征距离 $Dis\_E\ (A, B)$：

$$Dis\_E(A,B)=\sqrt{\sum_{i=1}^{12} w_{1\_i} * (A.Edgetype(i)-B.Edgetype(i))^2}$$

简单空间关系特征距离 $Dis\_R\ (A,B)$：

$$Dis\_R(A,B)=\sqrt{\sum_{i=1}^{12} w_{2\_i} * (A.Relation(i)-B.Relation(i))^2}$$

显而易见图形 A 和 B 越相似，则 $Dis_E\ (A,B)$和 $Dis_R\ (A,B)$值越小，所以实际上计算出的是相异度。为了得到真正的相似度且使得相似度值在[0,1]之间，令图形中共有 n 个边，我们做特征间归一化：

$$\begin{cases} Sim\_E(A,B)=\begin{cases} 0;\text{if } Dis\_E(A,B)>a*n; \\ 1-\dfrac{Dis\_E(A,B)}{a*n};\text{Else。} \end{cases} \\ Sim\_R(A,B)=\begin{cases} 0;\text{if } Dis\_R(A,B)>b*n; \\ 1-\dfrac{Dis\_R(A,B)}{b*n};\text{Else。} \end{cases} \end{cases}$$

其中 a、b 是常量，我们在实验中取2。这种变换实际上是设定某个相异度阈值（$a*n$ 和 $b*k$），当相异度超过这个阈值时，相似度就取0；这种归一化比较简单，且能保证相似度顺序保持不变。这样得到的相似度就能保证在[0,1]之间。且 A 和 B 相似度越高，则值越接近于1。

Step 3 计算总的相似度 Sim：$Sim(A,B)=W_1 * Sim\_E\ (A,B)+W_2 * Sim\_R(A,B)$。

Step 4 根据总相似度的降序提取 N 幅结果图形。结果图形的提取有两种策略：一是设定相似度阈值，返回所有相似度超过该阈值的图形；二是按相似度降序返回前 N 幅图形。在这里我们采取一个折衷策略：返回前 N 个结果中所有相似度小于阈值的结果图形（我们在实验中取相似度阈值为0.9；N 为8）。

Step 5 用户对结果图形集进行标注。对结果的标注分为三类：相关、中性、不相关。

Step 6 根据用户反馈调整权重。

Step 7 返回 Step 2，根据调整后的权重进行新一轮比较。

上述算法利用特征向量进行相似度匹配，简单直观、匹配速度快。假设标准图形库中标准图形个数为 k，对于算法空间消耗主要是存放标准图形库中图形的特征向量，又因为图形的特征向量是定长的，所以算法的空间复杂度和标准库中图形个数成正比，为 $O(k)$。对于时间复杂度主要是两个图形间的相似度比较，显然也和库中标准图形个数成正比，为 $O(k)$。

5 相关反馈

5.1 反馈层次

本文主要把握用户对图形整体形状上的意图（诸如旋转、形状的敏感度，如图5），不涉及底层的适应性和高层语义。图5中的3幅图形，有的用户认为(a)和(b)是最相似的，而有的用户则认为(a)和(c)是最相似的，我们在草图识别中引入相关反馈方法把握用户对图形总体形状的概念理解。

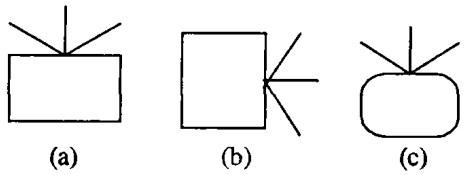


图5 用户意图举例

5.2 权重调整相关反馈

明确了反馈层次之后，我们需要一种反馈方法。我们的草图识别是以特征向量模型为基础，所以可以把一副图形看作特征空间中的一个向量点。权重调整方法的思想非常直观：每幅图形对应特征空间中的特征向量，识别开始时特征空间每一维度上的权重都是简单的初始值。权重调整方法把用户的意图映射到底层的特征权重上，通过调整权重来满足用户的需要，包括特征外部权重调整和特征内部权重调整。

·特征外部权重 W_1, W_2

设 RT 是初次识别中根据总相似度 Sim 得出的 n 幅最相似图形的集合，即作为候选识别结果返回给用户的 n 幅图形的集合：$RT=[RT_1, RT_2, \cdots, RT_n]$。如图6中左边为用户手绘草图，右边为按照相似度降序得到的8个候选识别结果。

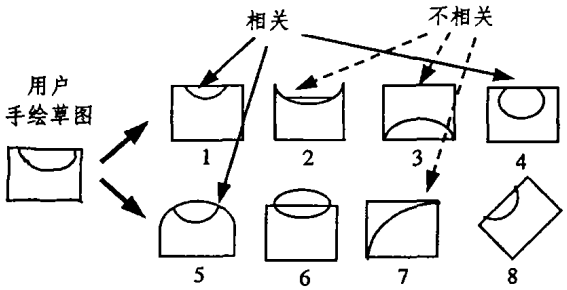


图6 识别结果和用户反馈

令 $SCORE_i$ 是用户对图形 RT_i 所给出的反馈得分：

$$SCORE(RT_i)=\begin{cases} 1;\text{如果图形是相关的；} \\ 0;\text{如果图形是中性的；} \\ -1;\text{如果图形是不相关的。} \end{cases}$$

在图6中，对于初次识别出的8幅图形，用户认为第1，4，5幅是相关的，第2，3，7幅是不相关的，而对第6，8幅无意见，也

就是认为它们是中性的。

同时只根据特征 *Edgetype* 的相似度 Sim_E 计算所得的最相似的 j 幅图像所组成的集合为：$RT^1=[RT_1^1, RT_2^1, \cdots, RT_j^1]$，同样只按照相似度 Sim_R 计算所得结果集合为 $RT^2=[RT_1^2, RT_2^2, \cdots, RT_k^2]$。接下来扫描 RT^1 和 RT^2 中每个元素就可以对权重 W_i 进行调整，对于元素 RT'_i（$i=1,2$）：

$$W_i=\begin{cases} W_i\times 1.2; & \text{If } RT'_i\in RT \,\&\&\, SCORE(RT'_i)=1;\\ W_i\times 0.95; & \text{If } RT'_i\in RT \,\&\&\, SCORE(RT'_i)=-1;\\ W_i; & \text{else。}\end{cases}$$

最后得到的权重 $W_i(i=1,2)$ 需要经过归一化：$W_i=W_i/(W_1+W_2)$。可以发现，如果 RT 和 $RT'(i=1,2)$ 集合中相关的图形重合越多（不相关的图形重合越少），相应的 W_i 值也越大，这也就是说，如果边类型特征或者简单空间关系特征反映了用户的信息需要，那么它将获得较大的重视程度。至于公式中的1.2和0.95是经验值，对于不同的反馈速度要求可以选取不同的值。

·特征内部权重 w_{i_j}

对于用户标注为相关的结果图形（总数为 M），在每个特征向量维度上应该存在 M 个值。从直观上讲，如果所有相关图像在该维度的值非常接近，那就意味着该维度很好地反映了用户的查询，应该给与较高的权重；相反，如果相关图像在某个分量上的值彼此相差很远，则说明该分量无法很好地反映用户的查询需求，则权重应该较小。根据以上分析，我们如下调整权重：

$$w_{i\_j}=1/(k+\delta_{i\_j})$$

其中 δ_{i_j} 为 w_{i_j} 对应维度上 M 个值的标准方差，k 用来平滑标准方差，我们在实验中取3。w_{i_j} 再经过和 W_i 一样的归一化操作后就得到调整后的特征内部权重。

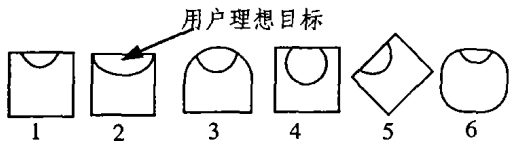


图7 最终识别结果

经过若干次反馈和权重调整后，权重逐渐逼近理想值，识别效果也越来越接近用户需求。对于图6中的识别，经过两次反馈后得到如图7中的6幅候选结果图形，此时结果2已经能够满足该用户的需要。

6 实验结果及分析

我们用实验来验证本文中识别方法的特点和有效性。实验是在 Intel P4 PC（2.0G Hz CPU，256MB 内存），Microsoft Windows XP 环境下进行的。我们设计了5组标准图形，每组有10个比较相似的图形共50个标准图形，另外加入了200个标准图形，标准图形库中共250幅标准图形。实验中由不同的用户对于每个手绘草图首先在标准图形库中指出他们认为相关的图形（至少一个），然后在此基础上分别用边表方法[2]、SRG方法[7]和本文中的方法进行识别，本文中方法对于每个识别进行5轮反馈。

为了有效评价实验结果，除了平均识别时间 *Time*(*ms*) 指标外，我们还借助信息检索中的相关反馈评价机制[8,9]：*Precision*、*Recall*、*Avg-r*、*Avg-p* 值。其中 *Precision* 为查准率，是候选集中相关图形数目与候选集图形总数的比例；*Recall* 为查全率，是候选集中相关图形数目与标准库中所有相关图形数目的比例。

对于 *Ave-r* 和 *Ave-p*：设 $Q_1, Q_2, \cdots Q_q$ 为输入草图，对第 i 个草图 Q_i，$I_1^{(i)}, I_2^{(i)}, \cdots, I_{a_i}^{(i)}$ 为相关的识别结果，$rank(I_j^{(i)})$ 为 $I_j^{(i)}$ 在识别结果集中的排序值，则可以用下列指标评价识别的效果：

$$Average\ r-measure=\frac{1}{q}\sum_{i=1}^{q}\frac{1}{a_i}\sum_{j=1}^{a_i}rank(I_j^{(i)});$$

这一指标定义了所有相关图形在识别结果集中的排序平均值。显然，该指标越小越好，说明相关图形在整体上的排列较为靠前。

$$Average\ p-measure=\frac{1}{q}\sum_{i=1}^{q}\frac{1}{a_i}\sum_{j=1}^{a_i}\frac{j}{rank(I_j^{(i)})};$$

这一指标定义了所有相关图形在识别结果集中靠前排列的紧密程度，如果所有图形均排在结果集的最前面，则这一指标等于1。显然该指标越大越好。

表1 草图识别实验结果

评价	边表方法	SRG 方法	引入相关反馈的草图识别方法					
			反馈前	反馈1次	反馈2次	反馈3次	反馈4次	反馈5次
Precision	0.380	0.432	0.352	0.430	0.458	0.479	0.499	0.500
Recall	0.710	0.720	0.772	0.772	0.805	0.852	0.893	0.935
Avg-r	5.460	4.950	5.975	5.117	4.583	4.375	4.091	3.816
Avg-p	0.504	0.528	0.577	0.638	0.671	0.696	0.716	0.740
Time(ms)	5.623	130.400	1.672					

表1显示了三种方法的实验结果。从表中可以看出：在反馈前，本文中方法的 Recall、Ave-p 和 Time 指标均优于其他两种方法，对于识别时间仅为1.672ms，本文方法具有较高的响应速度，特别适合在线处理。在用户反馈后，随着反馈次数的增加，各项指标均有提高，也就是说候选结果集中相关图形个数和排列均有提高。在反馈两次后所有指标均高于其他两种方法。但反馈达到一定次数后，效果提高逐渐趋于平缓。经过5次反馈后，已经能较好地把握用户意图，也就表明相关反馈取得了较好的效果。

总结 草图识别的模糊性和用户适应性是手绘草图识别领域中的一个难点问题。本文基于前期手绘草图识别系统的研究和开发，对上述问题进行了研究。以提取图形的特征向量为前提把相关反馈引入到草图识别过程中，与传统的草图识别方法相比，新的方法抽取出图形的向量化特征，大大降低了图形匹配的时间和存储空间。并通过相关反馈机制动态调整相似度匹配的特征内部权重和特征间权重，不断改进识别效果，最终准确捕获用户意图。文中所提出的方法取得了很好的实验效果。本文的研究只是在这一领域的一个初步尝试，我们

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)