一种基于EMD的文档语义相似性度量

王晓东[①②]　郭　雷[②]　方　俊[②]　董淑福[①]
[①](空军工程大学电讯工程学院　西安　710077)
[②](西北工业大学自动化学院　西安　710072)

摘　要：针对基于EMD(Earth Mover's Distance)的文档语义相似性算法不满足度量公理因而难以在信息检索与数据挖掘中推广应用的问题，该文提出了一种新的基于EMD的文档语义相似性度量——Mdss_EMD(Metric for document semantic similarity based EMD)。首先在分析EMD及现有改进方法缺陷的基础上，给出了文档宽度、虚拟项的概念；随后通过增加虚拟项来对齐文档矢量的总权值，使所有度量公理得到满足；最后，为提高该度量的适应能力及处理速度，还实现了虚拟项相似距离的弹性设计并对EMD算法进行了简化。该方法把EMD扩展到度量空间中来，很大程度上提高了EMD的索引能力与精度，初步实验表明，Mdss_EMD的整体性能优于原EMD及现有其它类似方法。

关键词：信息检索；EMD(Earth Mover's Distance)；度量；文档相似性；匹配；语义距离
中图分类号：TP391　　**文献标识码**：A　　**文章编号**：1009-5896(2008)09-2156-06

An EMD-Based Metric for Document Semantic Similarity

Wang Xiao-dong[①②]　Guo Lei[②]　Fang Jun[①]
[①](*The Telecommunication Engineering Institute, Air Foree Engineering University, Xi'an* 710071, *Chinese*)
[②](*College of Automation, Northwestern Polytechnical University, Xi'an* 710072, *China*)

Abstract: Aiming at the conflicts between EMD(Earth Mover's Distance)-based measure for document semantic similarity and metric axioms, which prevent EMD from being widely applied in the information retrieval and data mining, a novel EMD-based metric for document semantic similarity named Mdss_EMD is presented. Firstly, based on the analysis of drawbacks of EMD and its existing modifications, the concepts of document width and virtual term are proposed. Subsequently, by adding virtual term to initial document vector, the approach aligns the total weights of document vectors, so that all of metric axioms are satisfied. Finally, in order to improve the applicability and processing speed of the metric, the similarity distance of virtual term is designed to be elastic and EMD algorithm is also simplified. The proposed approach extends EMD to metric space, and substantially improves EMD on indexing and accuracy. The experimental results demonstrate that Mdss_EMD outperforms the original EMD and other similar measures in general.

Key words: Information retrieval; EMD(Earth Mover's Distance); Metric; Document similarity; Match; Semantic distance

1　引言

量化文档相似性的方法在信息检索和数据挖掘中具有十分重要的意义，它是文档分类、过滤、聚类、搜索等应用的基础计算，其性能优劣直接影响到信息检索和数据挖掘的质量与效果。文档相似性可以采用相似系数、相似距离等尺度进行衡量，本文将着重讨论基于相似距离的文档语义相似性度量(Metric)方法。

以往的文档相似距离，如：欧氏、海明距离等，一般认为文档特征元素/词汇相互正交而忽略了特征之间的语义关系，采用不同文档中相同词汇“一对一”的匹配方式进行比较计算，精度不够理想[1]。为了在计算中引入词汇语义关系并改进文档相似性计算中词汇匹配的方式，Wan等[1]利用图像检索领域中常用的EMD算法[2]及WordNet电子字典，实现了文档特征词汇之间“多对多”匹配的文档语义相似性算法，有效地提高了相似性计算精度。然而，EMD存在着只适合处理局部匹配、不满足度量公理的缺陷，使其在文档相似性计算的应用中受到了很大限制。针对这一问题，国内外学者提出了一些改进，如：Giannopoulos[3]提出PTD(Proportional Transportation Distance)函数、梁敏等[4]提出了X_dist柔性语义距离函数等，但这些改进后的算法在严格意义上讲，依然不能满足所有的度量公理要求，该问题已成为制约基于EMD的文档相似性算法性能提高与推广应用的主要矛盾。

受参考文献[3]的启发，本文在提出文档宽度和虚拟项等概念的基础上，通过增加虚拟项来对齐文档矢量的总权值，进而实现了基于EMD的文档语义相似性度量(Metric for document semantic similarity based EMD，Mdss_EMD)

2007-01-29 收到，2007-06-11 改回

方法。为提高度量的适应能力及处理速度，该方法还实现了虚拟项相似距离弹性设计并对原EMD算法进行了简化。Mdss_EMD把EMD扩展到度量空间中来，很大程度上提高了EMD的索引能力与精度，能够适合不同的应用环境。

本文其余部分安排如下：第2节介绍了基于EMD的文档语义相似性度量相关概念；第3节分析了EMD和现有改进的缺陷；第4节提出了文档宽度等概念和Mdss_EMD算法；第5节对Mdss_EMD进行了实验分析，最后一节给出本文结论和下一步工作。

2　相关概念

2.1 相似性度量模型

计算文档的相似性需要借助于相似系数或相似距离函数，二者可以互换。假设相似距离函数为 f，相似系数函数可以是 $1-f(\bullet)$ 或 $1/(1+f(\bullet))$ 等，这里 f 通常是一个度量，其定义如下：

定义1[3, 4]　一个度量空间是一个集合 S，连同一个函数 $\rho: S\times S\to r, r\geq 0$，使得任意的 $s_1, s_2, s_3\in S$，满足：

(1)自相似常数公理：$s_1=s_2, \rho(s_1,s_2)=0$。

(2)正性公理：$s_1\neq s_2, \rho(s_1,s_2)>0$。

(3)对称性公理：$\rho(s_1,s_2)=\rho(s_2,s_1)$。

(4)三角不等公理：$\rho(s_1,s_3)\leq\rho(s_1,s_2)+\rho(s_2,s_3)$。

则称函数 ρ 为 S 上的一个度量；如果仅能满足公理(1)，公理(3)，公理(4)，则称其为伪度量(Pseudo-metric)；如仅满足公理(1)，公理(2)，公理(3)，则称为半度量(Semi-metric)。

度量公理的约束对于度量应用具有非常重要的价值，特别是在信息检索与数据挖掘领域中，公理(2)对于辨识对象间特征差异，公理(4)对于使用三角不等式索引来提高检索效率，都有着至关重要的作用[3]。显然，伪度量和半度量因部分公理条件的缺失，在应用中将会受到严重限制。

2.2 词汇语义相似距离基函数

文档由词汇构成，在VSM(Vector Space Model)模型中通常把一个文档表示成为 n 维矢量 $\{(t_1,w_1),(t_2,w_2),\cdots,(t_N,w_N)\}$，特征词汇 t_k 称为项(可以是词组、短语、词等，一般取词)，w_k 为项 t_k 的 $\mathrm{tf}\cdot\mathrm{idf}$ 权重。计算文档之间的语义相似距离首先需要计算项之间的语义相似距离，相应的函数可记为 $D: w_i\times w_j\to R^+\bigcup\{0\}$，称之为语义相似距离基函数(以下简称基函数)。

目前计算词汇语义相似距离的主要方法就是，通过比较词汇在概念网络、领域本体(如：WordNet、HowNet等)或层次化结构中的路径长度、最小公共祖先、释义等加以实现[5, 6]。关于词汇语义相似距离基函数的设计不属于本文的讨论范围，这里不作详细介绍。

2.3　EMD 文档相似距离算法

虽然通过基函数可以很容易地求解项与项间的语义相似距离，但是在计算矢量空间模型中文档的相似距离时情况要复杂的多。这里不仅需要计算项间的语义距离，还应找到适合的项匹配。在以往的相似性算法中匹配因素通常被忽视，仅采用不同文档中的相同项“一对一”的匹配。实践证明这种方法在与人们的直觉贴近方面效果不佳，究其原因主要与词汇同义与多义现象有关。由于不同的文档可能采用不同的词汇来表示相同概念，上述“一对一”的匹配方法在处理时就显得无能为力了。针对这一问题，参考文献[1]将图像检索中常用的EMD算法引入到文档相似性计算中，提出利用EMD“多对多”的匹配特点对文档矢量项进行综合语义匹配，有效地提高了计算精度，下面给出文档的EMD相似距离的定义。

定义2[1]　设文档 $\boldsymbol{A}=\{(t_{a1},w_{a1}),(t_{a2},w_{a2}),\cdots,(t_{aN},\ w_{aN})\}$，$\boldsymbol{B}=\{(t_{b1},w_{b1}),(t_{b2},w_{b2}),\cdots,(t_{bN},w_{bN})\}$，有 $D=\{d_{ij}\}$，d_{ij} 为 t_{ai} 与 t_{bj} 的语义相似距离，$W=\sum_{i=1}^{N}w_{ai}$，$U=\sum_{j=1}^{N}w_{bj}$，另有匹配度 $F=\{f_{ij}\}$，f_{ij} 是由 w_{ai} 经 d_{ij} 匹配到 w_{bj} 的量，并满足：

$$f_{ij}\geq 0,\quad i=1,\cdots,N;\ j=1,\cdots,N \tag{1}$$

$$\sum_{j=1}^{N}f_{ij}\leq w_{ai},\quad i=1,\cdots,N \tag{2}$$

$$\sum_{i=1}^{N}f_{ij}\leq w_{bj},\quad j=1,\cdots,N \tag{3}$$

$$\sum_{i=1}^{N}\sum_{j=1}^{N}f_{ij}=\min(W,U) \tag{4}$$

现以 $\boldsymbol{A}$，$\boldsymbol{B}$ 的项集合 A，B 为两组顶点(下同)，连接两组顶点构成关系图 $G=\{A,B,D\}$，得到最小匹配总量 $\mathrm{Work}(A,B)$ 如下：

$$\mathrm{Work}(A,B)=\min_{f\in F}\sum_{i=1}^{N}\sum_{j=1}^{N}f_{ij}d_{ij} \tag{5}$$

则 $\boldsymbol{A}$，$\boldsymbol{B}$ 的相似距离定义为项集合 A 与 B 的EMD距离：

$$\mathrm{EMD}(A,B)=\frac{\mathrm{Work}(A,B)}{\min(W,U)} \tag{6}$$

在EMD计算中，可以把 $\boldsymbol{A}$ 的项看作质量分别为 w_{ai} 的若干堆土方，$\boldsymbol{B}$ 的项看作若干容量为 w_{bj} 的坑穴(反之亦然)，求文档 $\boldsymbol{A}$ 与 $\boldsymbol{B}$ 的相似距离问题即为求解将土方经距离为 d_{ij} 的路径填充到坑穴的最短距离运输方案，f_{ij} 为各路径上的流量，而 $\mathrm{EMD}(A,B)$ 即为最小的运输工作总量与相对轻的一方土方总质量或容量之比。在文档相似性计算中，EMD可以以灵活的“多对多”方式匹配同义词、多义词，甚至是近义词，从而提高计算的精度。

EMD通常可被直接描述为一个关于运输问题的线性规划。作为一类非常重要的优化问题，国内外众多学者已对该问题进行了大量深入的研究，并提供了许多EMD算法和工具[7, 8]，因而EMD在工程实践上也是十分可行的。

3　EMD 缺陷分析及现有改进

虽然EMD利用文档矢量 “多对多”的项匹配计算方法更加贴近于人们的直觉，但由于它不是专门为文档相似距离

计算设计的，所以在这方面的应用中存在一些不足，本节将对EMD及现有改进措施的缺陷进行分析。

3.1 EMD 主要缺陷

由定义2可以看出，EMD可以计算任意两个总权值不相等(即：$W \neq U$)的文档矢量之间的相似距离，多余的权值被忽略不计[3]，这就会导致EMD不符合度量的公理(2)、公理(4)，下面举例说明。

例 1 如图1所示，A，B，C，D 分别表示坐标轴上的4个点集，点集内各点之间的距离可以通过点所处的坐标位置得出。首先计算 $\mathrm{EMD}(A,B)$，其中 a_1 与 b_1 距离最近为1，a_2 与 b_2 距离最近为1，最小的运输工作量为 $2\times1+1\times1=3$，所以 $\mathrm{EMD}(A,B)=3/3=1$。同理，$\mathrm{EMD}(B,C)=0$，$\mathrm{EMD}(A,C)=8/6$，$\mathrm{EMD}(C,D)=0$。由此可知，$\mathrm{EMD}(A,C)>\mathrm{EMD}(A,B)+\mathrm{EMD}(B,C)$；$\mathrm{EMD}(C,D)=0$，$C\neq D$，这显然与定义1公理(2)和公理(4)不符。

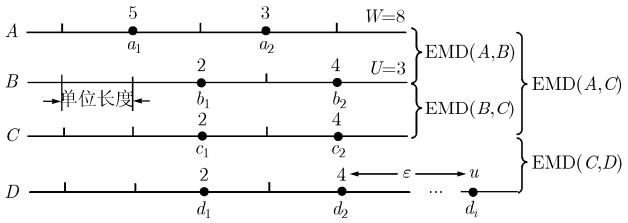

图1 EMD缺陷示意

3.2 PTD 改进方法

为了解决EMD的缺陷，Giannopoulos提出式(2)、式(3)取等号，在式(3)中给B的各项乘以比例变换系数 W/U，并用 $\sum_{i=1}^{N}\sum_{j=1}^{N}f_{ij}=W$ 替换式(4)使得运输总量以 W 为基准，从而实现了保留B各项之间的原始权值比例条件下的土方运输。PTD集合内的各项权值都被考虑进来，初步解决了与公理(4)相悖的问题[3]。

但是，PTD虽能够满足度量公理(1)、公理(3)、公理(4)，却不符合度量公理(2)，因而也仅是一种伪度量[3]。主要是因为PTD只考虑了集合各项的分布比例是否相等，而忽略权值总和上的差别，特别是当 $w_{ai}=W\cdot w_{bj}/U$ 时算法失效。在文档相似距离计算时，表现为分辨文档主题差异能力变弱。

3.3 X_dist 改进方法

文献[4]也提出了一种柔性语义距离X_dist对EMD进行改进，具体如下：

$$\mathrm{X\_Dist}(A,B)=\beta\cdot\mathrm{PTD}(A,B)+(1-\beta)\cdot\mathrm{X\_EMD}(A,B),\quad 0\le\beta\le1 \tag{7}$$

这里，X_EMD是EMD基于Jaccard系数的变形，相当于用 $\dfrac{\mathrm{Work}(A,B)}{W+U-\mathrm{Work}(A,B)}$ 或 $\dfrac{\mathrm{Work}(A,B)}{W+U}$ 替换式(6)。

实际上，X_dist也仅是X_EMD与PTD的线性组合，它利用两者优势进行互补从而初步解决了EMD的度量化问题。但是由于X_dist并不是严格从总权值恒等的理论角度改进的EMD(详见定理1)，因而不是真正意义的度量；并且X_dist要经过两次运输问题的求解，也严重降低了计算速度而不利于算法的推广。

4 Mdss_EMD 算法

本节详细介绍Mdss_EMD算法。

4.1 文档权重分配分析

定理 1[3, 9] EMD要完全满足度量公理需满足两个条件：①基函数本身为度量；②在EMD的计算空间域内，各集合总权值恒等(equal total weight sets)，即 $W\equiv U$。

在基于向量空间模型的文档相似性计算中，条件①容易实现然而②却很难达到，下面做简单分析。

表1列出了文档 tf·idf 项权值分配的主要模式，只需从3列因子中各取一种值就可以构成一个权值分配模式。不难看出：无论采用哪一种词频、文献频率和规范因子的组合来分配项权值，都不能保证文档矢量的总权值恒等。通常为了提高文档相似性计算的效率、节省计算空间，还要通过特征选择或抽取，降低矢量维数，将一部分权值较低的项直接或间接的从特征矢量中移除，这也加剧了权值的不平衡性。因而想要将EMD改进成为度量，必需解决权值不平衡的问题。

表1 文档项权值分配模式

词的频率因子		词的文献频率因子		规范因子	
代码	取值	代码	取值	代码	取值
b	1或0	n	1.0	n	1.0
n	tf	t	$\log(N/n)$	c	$1/\sqrt{\sum w_i^2}$
a	$0.5+0.5\cdot\mathrm{tf}/\mathrm{tf}_{\max}$	p	$\log((N-n)/n)$		
l	$\ln(\mathrm{tf})+1.0$				

4.2 Mdss_EMD

如前面3.2节对PTD的介绍，为了使EMD成为度量文献[3]对运输目标集 B 的权值进行比例变换，间接达到了平衡总权值的目的，但这种做法只考虑比例而未考虑权值总量差异，效果并不理想。受PTD启发，本文提出另一种新的总权值平衡方法，即：以项的形式为权值相对少的文档矢量补足权值差，然后对待计算相似距离的两个文档矢量进行归一化处理，再完成EMD计算，从而实现严格的文档语义相似性度量。

4.2.1 文档宽度与虚拟项

定义3 设 X 是一个集合，每一个映射 $M{:}x\to R^+\cup\{0\}$，$x\in X$，都称为分布 M 下 X 的一个分布值，$\sum_{x\in X}M(x)$ 是 X 在分布 M 下的宽度，记为 $\|X\|_M$，当 $X=\Phi$ 时，$\|X\|_M=0$。

在本文中，把项集合看作文档矢量的泛化，任何不区分项之间排列关系的文档矢量 $\boldsymbol{A}_i$ 都可简化表示为其项集合

A_i，因而当映射M取表1中任何一种 tf·idf 项权值分配模式时，$\|A_i\|_{\text{tf·idf}}$ 就表示为该模式下 A_i 的总权值，或称为文档 $\boldsymbol{A}_i$ 的宽度。两个相同项权值分配模式下的文档 $\boldsymbol{A}_i$ 与 $\boldsymbol{A}_j$ 的宽度之差称为文档宽度差，记为 W_{ij}，并有 $W_{ij}=\left|\|A_i\|_{\text{tf·idf}}-\|A_j\|_{\text{tf·idf}}\right|$。文档宽度差通常是降维处理、文档长度差异、项分布特征等多种因素引起的文档矢量总权值的差异，可以直观地解释为所有 t_k 对于不同文档的重要程度总体差异。这种差异显然应当被相似性计算考虑进来，然而在EMD，PTD，X_dist中并没有相应的计算步骤。

为此，在 EMD(A,B) 计算中，为宽度相对少的文档矢量以项形式补足文档宽度差(为了方便讨论，这里假设 $\|A\|_{\text{tf·idf}}\geq\|B\|_{\text{tf·idf}}$，以下均相同)，并称补项为虚拟项，记为：$(t_{b,N+1},w_{b,N+1})$，其权值等于 W_{AB}。之所以称之为虚拟项是因为它并非真实存在，而是由人工构造的文档间总权值差异的补充说明。利用虚拟项能够实现文档相似距离计算中总权值差异的引入。

至此，本文已经提出了对齐(平衡)EMD总权值的方法，为了使虚拟项能够参加EMD计算，除了对其赋权值外，还要给出该项与其它项的语义相似距离。虚拟项与其它各项的语义相似距离取值可以采用两种方法：

(1)与 t_{bj} 取值相一致，t_{bj} 满足 $\sum_{i=1}^{N}d_{ij}$ 为最大值，其中 t_{ai} 与 t_{bj} 对应的权值有 $w_{ai}\neq 0$，$w_{bj}\neq 0$；

(2)取均值 $\overline{d}$，$\overline{d}=\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}d_{ij}$，也有 $w_{ai}\neq 0$，$w_{bj}\neq 0$。

这两种方法都是对虚拟项语义特性的一种本地估计，与基函数的度量性质也并不矛盾。为了允许用户在应用中适当调整文档宽度差对相似距离的影响作用，提高了算法的适应能力，虚拟项的语义距离需乘以弹性系数 $1/(1-\gamma)$，$0\leq\gamma\leq 1$。

通过虚拟项的作用，满足了的 A，B 总权值相等的条件。为将该条件扩展到整个度量计算空间域中，算法在插入虚拟项后，对 $\boldsymbol{A}$，$\boldsymbol{B}$ 进行归一化处理，得到：$W'=\sum w_{ai}/\max(W,U)$，$U'=\sum w_{bi}/\max(W,U)$，并有 $W'\equiv U'=1$，从而在度量计算空间域内定理1要求的条件都能够得到满足，下面给出Mdss_EMD的定义。

4.2.2 Mdss_EMD 定义

定义4　设文档 $\boldsymbol{A}=\{(t_{a1},w_{a1}),(t_{a2},w_{a2}),\cdots,(t_{aN},w_{aN})\}$，$\boldsymbol{B}=\{(t_{b1},w_{b1}),(t_{b2},w_{b2}),\cdots,(t_{bN},w_{bN})\}$，$W=\sum_{i=1}^{N}w_{ai}$，$U=\sum_{j=1}^{N}w_{bj}$，$D=\{d_{ij}\}$ 为项相似距离集，$F=\{f_{ij}\}$ 是项匹配度集，不妨设 $W\geq U$，则插入虚拟项 $(t_{b,N+1},w_{b,N+1})$ 后 f_{ij} 满足：

$$\sum_{j=1}^{N+1}f_{ij}=w_{ai}/W,\quad i=1,\cdots,N \tag{8}$$

$$\sum_{i=1}^{N}f_{ij}=w_{bj}/W,\quad j=1,\cdots,N+1 \tag{9}$$

$$\sum_{i=1}^{N}\sum_{j=1}^{N+1}f_{ij}=1 \tag{10}$$

则 $\boldsymbol{A}$，$\boldsymbol{B}$ 的相似距离定义为项集合 A 与 B 的Mdss_EMD相似距离：

$$\text{Mdss\_EMD}(A,B)=\min_{f\in F}\sum_{i=1}^{N}\sum_{j=1}^{N+1}f_{ij}d_{ij} \tag{11}$$

性质 1　给定n维空间下的任意3个文档矢量 $\boldsymbol{A}$，$\boldsymbol{B}$ 和 $\boldsymbol{C}$，对应的项集合为 A，B 和 C，基函数为度量的条件下，Mdss_EMD有以下性质：

(1) A 与 B 语义分布相等时，Mdss_EMD$(A,B)=0$。

(2) A 与 B 语义分布不等时，Mdss_EMD$(A,B)>0$。

(3) Mdss_EMD(A,B) = Mdss_EMD(B,A)。

(4) Mdss_EMD$(A,C)\leq$Mdss_EMD(A,B) +Mdss_EMD(B,C)

需说明的是，由于Mdss_EMD是基于语义的相似距离，所以性质(1)、性质(2)应当在语义条件下进行讨论，例如：由于 D(transport,conveyance)$=0$，所以认为{(transport, 0),(conveyance,0.1),(produce,0.1)} 与 {(transport,0.1), (conveyance,0),(produce,0.1)} 相等，这符合文档语义相似性计算实际情况。

4.3 算法描述

为了进一步提高运行速度，对Mdss_EMD的EMD算法部分进行简化，简化过程遵循语义相似距离最小的项优先匹配原则，具体过程如下：

(1)取文档矢量 $\boldsymbol{A}$，$\boldsymbol{B}$ 的项集合 A，B，令 $D=\{\}$，$F=\{\}$，$\|A\|_{\text{tf·idf}}$，$\|B\|_{\text{tf·idf}}$，$W_{\boldsymbol{A},\boldsymbol{B}}$ 置0；

(2)计算 $D=\{d_{ij}\}$，$\|A\|_{\text{tf·idf}}$，$\|B\|_{\text{tf·idf}}$，$W_{\boldsymbol{A},\boldsymbol{B}}$，$d_{i,N+1}$；

(3)插入虚拟项并更新 D；

(4)新文档矢量归一化处理；

(5)选择最小的 d_{ij}，记录其对应的 (i,j)；

(6)由 d_{ij} 和 (i,j) 计算 $\min(w_{ai},w_{bj})\rightarrow f_{ij}$；

(7)如果 $w_{ai}=0$ 或 $w_{bj}=0$，则执行下一步；如果 $w_{ai}\leq w_{bj}$，则 $w_{bj}=w_{bj}-w_{ai},w_{ai}=0$；如果 $w_{ai}>w_{bj}$ 则 $w_{ai}=w_{ai}-w_{bj},w_{bj}=0$；

(8) $D-\{d_{ij}\}\rightarrow D$；

(9)如果 $\sum w_{ai}\neq 0$，返回第5步；

(10)按照公式(11)计算 Mdss_EMD(A,B)。

5 实验

5.1 实验环境

实验软硬件环境：P43.0Ghz CPU，内存512M，硬盘80G；Windows XP Professional操作系统，NTFS文件系统；实验

语料选取3942篇来自于Reuters-21578的文档，主要选择其中包含文档最多的10个类中，文档长度适中、主题内容较集中、具有人工标注类别的文档。

程序设计关键部分实现方法如下：首先对测试文档进行矢量化处理，包括停用词去除、词根还原、项权值分配，分配采用“ltc”模式；将得到的文档矢量以文本文件的方式存储；再利用参考文献[10]提供的Perl工具“lesk”算法完成项语义距离计算，虚拟项相似距离赋值采用4.2.1节方法(2)，存储计算结果；主程序采用VC++ 6.0实现，其中EMD、PTD、X_dist函数(以X1_dist为代表)计算部分利用参考文献[8]所提供的EMD源码修改得到，Mdss_EMD采用4.3节的简化设计。

5.2 文档分类应用

本实验的目的是考察几种相似距离在信息检索与数据挖掘常用的文档分类中的精度。测试方法：在10类文档中每类取35篇共计350篇文档作为测试文档，其余部分作为训练文档，分别使用Euclidean、EMD，X1_dist，PTD，Mdss_EMD对kNN进行程序实现，然后进行不同k值下的分类测试。分别取 $k=1,5,10,20,30$ ，与 k 对应的kNN分类阀值 $b=2,11,24,66,81$ ， $\gamma=0.5$ ， X1_dist调节参数为0.8，采用式(12)所示的评估方法：

$$F_\beta(r,p)=\frac{(\beta^2+1)pr}{\beta^2 p+r} \tag{12}$$

其中 r 为召回律， p 为准确率， β 为指标，取 $\beta=1$ 。

实验结果如图2所示，可以看出Euclidean精度最差，Mdss_EMD优于EMD，PTD介于二者之间，X1_dist精度最好。上述实验结果首先证实了“多对多”的项匹配算法精度均要好于“一对一”的算法；其次表明Mdss_EMD的精度要优于EMD和PTD却低于X1_dist，这主要是因为EMD和PTD在召回率相当的情况下由于分辨能力弱于Mdss_EMD导致准确率 p 偏低，致使整体精度低于Mdss_EMD，而X1_dist采用线性组合的方法同时继承了EMD和PTD的优点，因而性能较Mdss_EMD稍好，但这是在牺牲计算效率的前提下达到的。

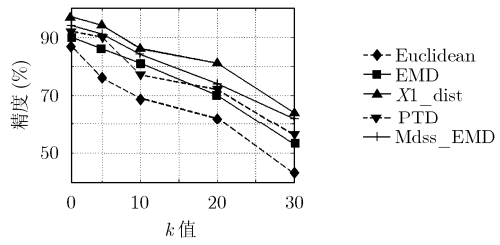

图2 分类精度比较

5.3 计算时间

本实验目的主要是比较几种相似距离的计算时间。由于几种算法在文档预处理、特征词提取、矢量表示、项语义距离计算部分(除去Euclidean之外)的工作基本相同，因而这部分消耗的计算时间不考虑在内。测试方法：从上述Reuters-21578的测试集中随机抽取完全不同的两篇文档50对，分别由各算法计算相似距离，记录总运行时间后求平均，计算分别在维度为50，100，200，300，400，500的条件下进行。

实验结果如图3所示，可知Euclidean算法速度最快，经简化计算的Mdss_EMD次之，其它依次是EMD，PTD，X1_dist。原因分析：Euclidean算法最简单；EMD、PTD在算法上差别不大，但是EMD在完成最小权值总量的运输后就结束计算，而PTD要对所有权值进行运输所以计算时间稍长；X1_dist相当于进行了EMD与PTD各一次，时耗最长；Mdss_EMD不但只进行一次运输计算，而且是简化处理后的EMD算法，因而速度很快，仅次于Euclidean算法，非常接近于实用。

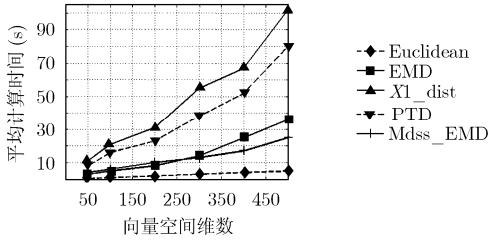

图 3 计算时间比较

由上述实验可得出以下结论：Mdss_EMD继承了EMD“多对多”的匹配特点，且考虑了权值总和的差异，是一种严格的度量，精度高、速度快，具有良好的推广潜力。

6 结束语

本文提出了一种新的基于EMD的文档相似性度量方法Mdss_EMD，该方法继承了EMD“多对多”的匹配特点，且考虑了权值总和的差异，可以作为一种严格的度量在面向文档的信息检索与数据挖掘中推广应用。本文所做研究对于丰富和完善文档相似性量化方法做出了贡献。

未来的工作可以从下面几个方面展开：首先，在信息检索和数据挖掘的多种应用中进一步对Mdss_EMD进行测试和评价；其次，根据实验结果的指导，提出适合不同应用的虚拟项与别项的语义距离赋值、以及 β 调节方法；再有，将虚拟项的概念推广到集合相似性度量中去，产生更多具有实用价值的相似性度量方法。总之完善Mdss_EMD的研究任务还很多，也非常具有研究意义。

参 考 文 献

[1] Wan Xiaojun and Peng Yuxin. The earth mover's distance as a semantic measure for document similarity[C]. ACM Fourteenth Conference on Information and Knowledge Management (CIKM), Bremen, 2005: 301-302.

[2] Rubner Y and Carlo T, *et al*.. The Earth mover's distance as

a metric for image retrieval[J]. *International Journal of Computer Vision*, 2000, 40(2): 99–121.

[3] Giannopoulos P and Remeo C V. A pseudo-metric for weighted point sets[C]. The 7th European Conf on Computer Vision, Copenhagen, 2002: 715–730.

[4] 梁敏, 郭新涛, 等. X_dist—一个柔性语义距离函数[J]. 计算机研究与发展, 2004, 41(10): 1728–1736.
Liang Min and Guo Xin-tao, *et al.*. X_dist——a flexible semantic distance function[J]. *Journal of Computer Research and Development*, 2004, 41(10): 1728–1736.

[5] Pedersen T and Patwardhan S, *et al.*. WordNet: Similarity-measuring the relatedness of concepts[C]. In Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics Demonstrations, Boston, 2004: 38–41.

[6] Prasanrua G and Hector G M, *et al.*. Exploiting hierarchical domain structure to compute similarity[J]. *ACM Trans on Information System*, 2003, 21(1): 64–93.

[7] Svetlozar T R and Ludger R. Mass transportation problems[M]. Volume I: Theory, New York : Springer-Verlag, 1998: 36–180.

[8] Rubner Y. Source code for the EMD software[CP]. http: // robotics. stanford.edu/~rubner/emd/default.htm, Retrieved 2007, 1.

[9] Yossi Rubner. Perceptual metrics for image database navigation[D]. Stanford University, Department of Computer Science, 1999.

[10] Pedersen T and Patwardhan S, *et al.*. WordNet::similarity-perl modules for computing measures of semantic relatedness [CP]. http://search.cpan.org/dist/WordNet-Similarity/lib/WordNet/Similarity.pm, Retrieved 2007, 1.

王晓东： 男，1974年生，博士生，研究方向为信息检索、本体、数据可视化.

郭 雷： 男，1956年生，教授，博士生导师，研究方向为神经计算、计算机视觉、图像处理等.

方 俊： 男，1981年生，博士生，研究方向为信息检索、本体、语义网.