北京大学学报(自然科学版)
Acta Scientiarum Naturalium Universitatis Pekinensis

基于排序学习的文本概念标注方法研究

涂新辉[1,2,†]　何婷婷[1,2]　李芳[1,2]　王建文[1,2]

1. 华中师范大学计算机学院，武汉　430079; 2. 国家语言资源监测与研究中心网络媒体语言分中心，武汉　430079; †E-mai: tuxinhui@gmail.com

摘要　提出一种基于排序学习的方法来实现文档的维基百科概念自动标注。首先人工对一定规模的文档进行概念标注，建立训练集合，然后利用排序学习算法在多项特征上得到对概念排序的模型，利用这个概念的排序模型对任意文档进行概念标注。实验表明，相对于传统的文档概念标注方法，此方法在各类指标上都有相当大的提高，标注结果更加接近人类的概念标注。

关键词　概念标注；排序学习；维基百科；显示语义分析

中图分类号　TP391

Learning to Annotate Text Using Wikipedia Concepts

TU Xinhui[1,2,†], HE Tingting[1,2], LI Fang[1,2], WANG Jianwen[1,2]

1. School of Computer Science, Huazhong Normal University, Wuhan 430079; 2. Network Media Branch, National Language Resources Monitoring and Research Center, Wuhan 430079; † E-mail: tuxinhui@gmail.com

Abstract　This paper proposed an automatic text annotation method based on learning to ranking model. Firstly the authors built a training set of concept annotation manualy, and then used the Ranking SVM algorithm to generate concept ranking model, finally the concept ranking model was used to generate concept annotation for any texts. Experiments show that proposed method has a significant improvement in various indicators compared to traditional annotation methods, and concept annotation results is closer to human annotation.

Key words　concept annotation; learning to ranking; Wikipedia; explicit semantic analysis

人类理解自然语言的过程是一个语义概念的联想和关联的过程，这种功能是由人类大脑中几百亿个神经元构成的复杂生理组织所提供的。建立基于概念的文本表征模型是实现基于语义的文本内容处理的一个途径[1]。能够表征文本中所蕴涵的各种复杂主题所使用概念集合应该满足以下条件：1) 包含覆盖不同领域主题的海量概念；2) 新的概念能够及时加入到这个概念库中；3) 这些概念应该是人可以理解的。要建立和维护这样应该自然概念的集合是一个异常艰巨的任务。幸运的是，维基百科——这个世界上最大的百科知识库——已经满足了上面的几个要求。

已有的维基百科概念自动标注方法可以分为两类：基于关键词匹配和基于内容匹配。前者主要通过文本中出现的词语或词组来标示出概念；而后者则通过文档和概念的相关度来进行概念的标注。

最早的基于关键词匹配的维基百科概念标注方法是由文献[2]提出的 Wikify 系统。这个系统的概念标注过程可以分为探测和消歧两个部分。首先，在文本中探测可能标注为概念的词语或词组，在链接到维基百科概念时利用了链接概率的算法，基本思想是统计词组在维基百科文本中链接到维基百科概念的概率，在文本中分析所有可能的 N 元词串并计算其链接概率，链接概率较大的将被确认为候选链接。然后，对链接的候选概念进行消歧，确保链接到正确的概念。对于文本中的大部分锚文本，可

国家自然科学基金(90920005, 61003192)资助
收稿日期: 2012−05−31；修回日期: 2012−08−13；网络出版时间：2012-10-26 17:55
网络出版地址：http://www.cnki.net/kcms/detail/11.2442.N.20121026.1755.025.html

能链接到多个不同的概念。例如: “plane”这个词语大多数情况下链接到飞行器这个概念, 但是在有些情况下可能链接到空间中的超平面的概念, 或者链接到木材加工时使用的刨子。为了选择正确的概念, Wikify 利用锚文本的上下文中出现的词语作为特征, 利用维基百科语料为训练数据, 得到各个概念的分类器, 这些分类器将对文本中的锚文本链接到的候选概念进行分类判别, 选择正确的概念。Milne 等[3]提出了一种改进的候选概念消歧的方法, 在对候选概念进行消歧时, 引入了概念的通用度作为一个参考依据, 大大提高了消歧的精度。

Maron[4]提出了一种称为话题索引的方法, 这种方法的目的是标示出系统中主要话题。这些识别出的话题可以被用来标注和组织文档以及概括文档的核心思想。这种方法和 Wikification 系统十分类似, 主要的区别在于对概念的选择。Medelyan 等[5]也提出了类似的话题索引的方法。

Ferragina 等[6] 提出了一种高效的基于关键词匹配的概念标注方法, 在不影响精度的情况下大大提高了标注的效率。Kulkarni 等[7]提出了一种改进的标注方法, 在选择概念时不仅考虑概念的局部信息, 同时还考虑了概念之间的一致性等全局信息。

基于内容匹配的方法中影响最大的是显示语义分析 (explicit semantic analysis)[8]。与基于关键词匹配方法不同的是, 显示语义分析方法直接利用文本和维基百科概念对应的文档之间的基于词语向量的相似度对概念进行排序, 相关度最高的概念将被选择作为表征文档主题的概念集合。

通过基于关键词匹配的方法得到的概念往往不一定能够很好地表示文本的主题信息, 而已有的基于内容匹配的标注方法存在一个共同的问题: 只考虑了文本内容和维基百科概念的相关性, 而没有考虑维基百科概念的本身的特征等其它方面的因素。

本文提出了一种基于排序学习的方法来实现文档的维基百科概念自动标注。首先人工对一定规模的文档进行概念标注建立训练集合, 利用 Ranking SVM 排序学习算法得到对概念排序的模型, 利用这个概念的排序模型对任意的文档进行概念标注。实验表明, 本文中的方法比传统的文档概念映射方法在各类指标上都有相当大的提高, 这个标注系统比传统的方法得到的概念标注更加接近人类的概念标注。

1 基于排序学习的概念标注方法

本文中概念标注方法包括两个部分: 1) 维基百科概念库的构建, 包括维基百科中概念的获取和筛选; 2) 文本的自动概念标注, 利用基于机器学习的方法实现文本的概念自动标注。

1.1 维基百科概念库的构建

维基百科诞生于2001年, 目前是世界上规模最大的百科全书。其英文版包含的条目数量是大英百科全书的10倍以上, 而且英文版只是维基百科包含的 250 个不同语言版本的一部分。维基百科上的页面由互联网上近 10 万个正规的贡献者编辑和维护。维基百科上的信息具有很强的实时性, 往往新发生的事件在几天内就被添加为维基百科条目。

在维基百科中, 信息是按照一套预先定义的结构来组织和创建的。主要的结构元素包括: 主题页面、重定向页面、消歧页面和类别等。

1) 主题页面。主题页面是维基百科中最重要的元素。每一个主题页面都代表一个单独的概念, 其标题是一个严格定义、具有格式统一的词语或词组。在维基百科中的标题是其唯一的标示符, 歧义页面通常使用附加的信息加以区分。例如: “Mouse”这个标题代表动物老鼠这个概念的页面, 而“Mouse_(computing)”这个标题则是计算机鼠标页面的标题。页面中通常使用不同语种的自然语言来描述这个概念的相关信息, 页面中的信息都是和这个概念密切相关的, 可以被看做为这个概念的语义上下文。

2) 重定向页面。在自然语言中, 存在很多同义现象, 即多个词语表达相同的概念。在维基百科中, 如果多个概念是等同的, 那么这些概念中除了一个概念的页面中包含概念的描述以外, 其它的概念使用重定向的链接映射到这个页面中。这种这包含重定向链接的页面被称为重定向页面, 这种方式避免了概念的重复定义, 同义的概念被组织一个共同的信息页面, 在一定程度上也简化了信息的维护。这种重定向页面的机制还被用于处理大写方式、拼写变体、缩写以及专业术语等问题。例如: 在英文维基百科中, 有一个页面为“United States”, 指向它的重定向页面包括“U.S.”, “USA”, “US”。

3) 消歧页面。和同义现象相反, 自然语义中还普遍存在歧义的现象, 即一个词语可以表达多个不同的概念。在维基百科中, 消歧页面就是专门处理

歧义现象的一种机制。例如: 在“Apple_(disambiguation)”这个消歧页面中, 存在多个不同链接到以下多个概念的页面: “Apple”, “Apple_Inc”等。

4) 类别。类别是维基百科中对概念页面信息进行组织的一种有效的手段。通常, 每一个主题页面至少归属于一个类别。例如: “China”这个主题页面归属于“East Asia”等多个类别。类别本身不是专题页面, 它们的存在只是为了便于组织和管理页面。类别的目标是建立信息的层次关系, 实际上维基百科的类别并不是严格的树型结构, 而是一种接近图形的复杂结构。

维基百科是一个便于人阅读和编辑的知识库, 但是计算机并不能直接利用这个知识库系统中的语义信息。而且, 作为一个开放的人人都可以编辑的知识库, 维基百科中不可避免地存在很多不准确的信息和噪音信息。为了得到一个干净简洁的计算机可以进行处理的概念库, 我们必须进行进一步的加工预处理。

每一篇维基百科文章的标题通常可以对应为一个概念。但是, 有一些专门用于维基百科信息管理的文章标题是没有意义的, 例如: “1980 年代”, “新闻报刊列表”等。这里, 我们根据以下原则对这些信息进行筛选, 满足以下任意条件的将被去取: 1) 标题和时间信息有关的文章, 例如: “年代”, “世纪”等; 2) 标题的首字母没有被大写的文章; 3) 标题是停用词的文章; 4) 对于一个包含多个单词的标题, 除去代词、连接词、限定词的其他某个词语没有大写的文章; 5) 文章的标题在文章中出现的次数少于 3 次。

按照上面的规则对主题页面进行筛选后得到符合要求的概念集合。在维基百科文档集中, 文档之间存在大量链接, 链接中的锚文本是往往是概念名称的一个重要的信息源。通过分析整个维基百科文档集合, 可以得到概念的名称、维基百科概念、主题页面及其映射关系的信息, 如图 1 所示。

1.2 自动标注方法

通过传统的文本相关性计算模型可以得到与文档相关的概念, 但是影响文档和概念相关性因素是多方面的, 而且有一些概念并不是很适合表达文档的主题。人工建立的计算模型很难综合考虑各种特征的影响, 本文使用了基于排序学习的方法得到相关性计算的模型。排序学习是一种生成排序模型的重要方法, 相关研究人员已经提出了各种不同的排序学习的方法[9]。这种方法的基本思想为: 首先建立训练集合, 人工为文档标注维基百科概念, 然后利用排序学习算法得到排序模型。

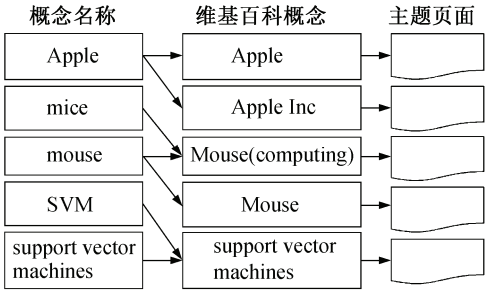

图 1 概念库的结构

Fig. 1 Framework of the concept collection

本节中我们使用经典的 Ranking SVM[10]方法。Ranking SVM 的输入是针对一组查询的偏序排序信息的一组训练集合:

$$(d_1, r_1), (d_2, r_2),\ldots,(d_n, r_n) \quad (1)$$

d_1 是一个文档, r_1 是和文档相关的概念的相关性级别信息。如果相对于文档 d_k, 概念 c_i 比 c_j 的相关程度高, 那么$(c_i, c_j)\in r_k$, 否则$(c_i, c_j)\notin r_k$。这些排序的信息可以从通过人工判定得到的训练数据中得到。

假设排序学习方法得到一个线性的排序函数 $\boldsymbol{w}\cdot\boldsymbol{c}_a$, 其中 $\boldsymbol{w}$ 为一个通过学习得到的权重向量, $\boldsymbol{c}_a$ 为概念 $\boldsymbol{c}_a$ 的特征向量, 这些特征可以通过各种不同的角度获取, $\boldsymbol{w}$ 决定了每个特征的重要性。如果一个概念的特征向量表示为 $\boldsymbol{c}_a$ =(3, 5, 1), 权重向量为 $\boldsymbol{w}$=(2, 3, 2), 那么计算排序函数的分值就是

$$\boldsymbol{w}\cdot\boldsymbol{c}_a=(2, 3, 2)\cdot(3, 5, 1)=6+15+2=23。\quad (2)$$

给定概念的训练数据和相关性排序信息, 我们希望找到一个权重向量尽可能多地满足以下条件:

$$\begin{aligned}&\forall (c_a, c_b)\in r_1: \boldsymbol{w}\cdot\boldsymbol{c}_a>\boldsymbol{w}\cdot\boldsymbol{c}_b,\\ &\forall (c_a, c_b)\in r_n: \boldsymbol{w}\cdot\boldsymbol{c}_a>\boldsymbol{w}\cdot\boldsymbol{c}_b,\end{aligned} \quad (3)$$

也就是说, 对排序数据中的所有概念对, 希望更适合表示文档主题的概念在排名上超过不适合表示文档主题的概念。实际上, 没有有效的方法找到完全满足条件的权重向量 $\boldsymbol{w}$, 但是, 可以把这个问题转换为一个标准的 SVM 优化问题:

$$\begin{aligned}&\mathrm{Min}:\frac{1}{2}\boldsymbol{w}\cdot\boldsymbol{w}+c\sum\xi_{a,b,k},\\ &\forall (c_a, c_b)\in r_1: \boldsymbol{w}\cdot\boldsymbol{c}_a>\boldsymbol{w}\cdot\boldsymbol{c}_b+1-\xi_{a,b,1},\\ &\forall (c_a, c_b)\in r_n: \boldsymbol{w}\cdot\boldsymbol{c}_a>\boldsymbol{w}\cdot\boldsymbol{c}_b+1-\xi_{a,b,n},\\ &\forall_a\forall_b\forall_k:\xi_{a,b,k}>0,\end{aligned} \quad (4)$$

其中ξ为松弛变量, 用于控制训练数据中噪音的影响, C 是防止过拟合的参数。过拟合是指学习算法在训练数据上表现完美, 但是无法适应新的数据。

在评估一个概念是否适合表达文档的主题时，需要考虑多方面的因素，包括概念的一些基本特征与文档和概念的内容相关性等。实验中我们使用的特征如表 1 所示。

我们使用语义模型中基于 KL 交叉熵的方法来计算文档 d 和概念 c 的相关性。计算公式如下:

$$\begin{aligned}\mathrm{SIM} = (C, D) &= -KL(\theta_D \,\|\, \theta_C) \\ &= -\sum_{t\in V} P(t\,|\,\theta_C)\log\frac{P(t\,|\,\theta_C)}{P(t\,|\,\theta_D)} \\ &= -\sum_{t\in V} P(t\,|\,\theta_C)\log P(t\,|\,\theta_D) - \\ &\quad \sum_{t\in V} P(t\,|\,\theta_C)\log P(t\,|\,\theta_C), \end{aligned} \tag{5}$$

公式中的 θ_D 和 θ_C 分别为文档模型和概念模型，$P(t|\theta_D)$和 $P(t|\theta_C)$分别是文档和概念相对应的维基百科文档中词语 t 的生成概率。

2 实验配置

2.1 数据预处理

作为一个开源的项目，维基百科机构每隔几天或几周都会把所有的文档数据以数据库文件的方法发布到网站上，供免费下载。我们的实验使用的维基百科的版本为 2008 年 5 月 24 日发布的。包括修改日志在内的全部数据压缩后大小为 120 GB。我们只需要使用其中 2.3 GB 的包含基本文档信息的数据。这个版本的英文维基百科共包含约 300 多万篇文档，经过上面介绍的方法筛选后得到共 531439 个概念的集合。

表 1　各类特征列表

Table 1　List of various types of features

序号	特征名	描述
1	INLINK(c)	概念所对应维基百科文档的入链接数
2	OUTLINK(c)	概念所对应维基百科文档的出链接数
3	CAT(c)	概念所对应维基百科文档的类别数
4	REDIRECT(c)	链接到这个概念的重定向页面数量
5	LEN(c_body)	概念对应的维基百科文档正文长度
6	LEN(c_title)	概念对应的维基百科文档标题长度
7	C_IN_D(c, d)	概念的语义标签是否在文档中出现
8	TF(c, d)	文档中包含概念的语义标签的频次
9	SC_IN_D(c, d)	概念语义标签的子串是否在文档中出现
10	SIM (d, c)	文档 d 和概念 c 的内容相关性

2.2 训练样本构建

在本文的实验中，12 位用户使用一个专门开发的文档段落标注系统，为 600 个段落进行了人工标注，系统的界面如图 2 所示。这些标注的数据作为分类器的训练数据。在标注过程中，首先为每篇文档列出了候选的维基百科概念列表，用户根据概念和段落主题的相关性以及概念本身是否适合表达主题等信息进行人工判定，最后选取不超过 10 个概念作为段落的概念表征。标注完成后共得到 3387 个概念，平均每个段落 5.6 个概念。在实验中我们同时选取了未被选择的 3500 个概念作为反例。

2.3 训练样本分组

为了进行交叉验证，我们把训练数据划分为 5 个子集，分别表示为 S1, S2, S3, S4 和 S5。在每一组实验中，使用其中 3 个子集作为训练集，剩下的两个子集分别作为验证集和测试集。训练集用于学习排序模型，验证集用于调整模型的参数，如 Ranking SVM 中目标函数的联合系数，测试集用于报告模型的效能。需要注意的是因为使用了 5 组交叉验证方法，本文中的模型的效能实际上是 5 组实验的平均值。

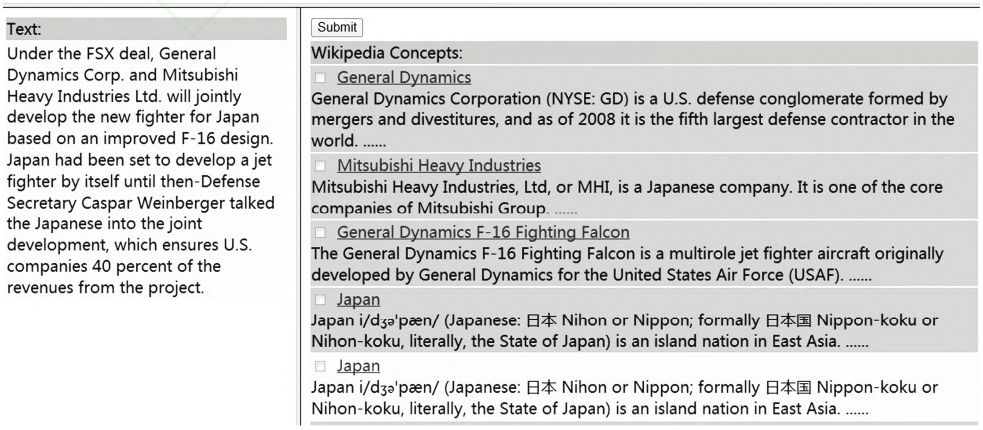

图 2　段落相关概念人工标注系统

Fig. 2　Relevant Wikipedia concepts confirmation system

2.4 特征归一化处理

由于不同文档之间特征的绝对值可能是不可比较的, 在本文的实验中, 我们对每一个特征进行了归一化处理。假设和文档 d_i 相对应的概念为 $\{c_j^{(i)} | j = 1,\ldots,N^{(i)}\}$, 概念 $c_j^{(i)}$ 的特征表示为 $x_j^{(i)}$ ($j = 1,\ldots, N^{(i)}$), 则归一化以后, 特征的值可以通过以下公式计算:

$$\frac{x_j^{(i)} - \min\{x_k^i, k=1,...,N^{(i)}\}}{\max\{x_k^{(i)}, k=1,...,N^{(i)}\} - \min\{x_k^{(i)}, k=1,...,N^{(i)}\}}。(6)$$

3 实验结果分析

3.1 不同概念数的比较

为了检验排序学习模型的稳定性, 我们测试了不同概念数情况下模型的精确度。精确度的计算是基于 5 组实验的平均值, 精确度在概念数变化时的变化情况如图 3 所示。可以看出, 在概念数增加时精确度逐步下降。本文中的基于 Ranking SVM 算法的排序学习模型和基准方法使用 CRM(concept ranking model) 表示。

3.2 与已有方法的比较

为了检验排序学习模型的性能, 我们以基于内容相关度的方法作为基准方法, 使用以 SIM 表示。显示语义分析 (ESA) 是利用词和概念共现的信息, 来实现文本到维基百科概念映射的技术, 在文本分类、聚类和信息检索领域已经取得了一定的成功。为了进一步检验本文中 CRM 模型的效能, 我们同时比较了 ESA 和 CRM 模型在测试集合上的表现。比较结果如表 2 所示。

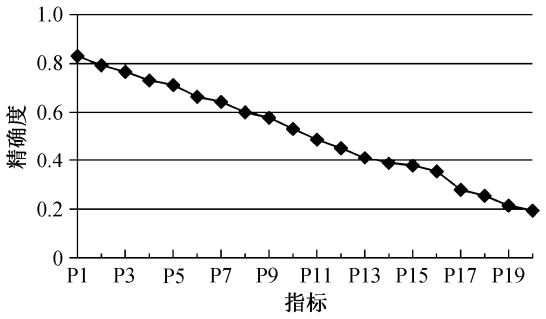

图 3 概念数增加时精确度的变化情况

Fig. 3 Change of the precision with the increase of the number of concepts

表 3 给出了对例句“Olympic News In Brief: Cycling win for Estonia. Erika Salumae won Estonia’s first Olympic gold when retaining the women’s cycling individual sprint title she won four years ago in Seoul as a Soviet athlete.”的概念标注。从表 3 中的数据可以看出, 相对于传统的基于内容相关度的方法, 通过排序学习模型得到的概念能够更好地表

表 2 实验结果比较

Table 2 Comparison of experimental results

评价指标	SIM	ESA	CRM	相对于 SIM 提高百分比/%	相对于 ESA 提高百分比/%
P1	0.5492	0.6732	0.8310	51.3	23.4
P5	0.4353	0.5318	0.7125	63.7	34.0
P10	0.3419	0.4165	0.5323	55.7	27.8
P15	0.2355	0.3401	0.3781	60.6	11.2
P20	0.1136	0.1527	0.1934	70.2	26.7
MAP	0.4082	0.5013	0.6215	52.3	24.0
Recall	0.6145	0.6942	0.7853	27.8	13.1

表 3 生成的概念标注示例

Table 3 Examples of generated concepts

序号	ESA	CRM
1	Estonia at the 2000 Summer Olympics	Erika_Salumäe
2	Estonia at the 2004 Summer Olympics	Olympic Games
3	2006 Commonwealth Games	Sport_in_Estonia
4	Estonia at the 2006 Winter Olympics	Summer Olympic Games
5	1992 Summer Olympics	Estonia at the Olympics
6	Athletics at the 2004 Summer Olympics Women’s Marathon	Olympic Gold
7	2000 Summer Olympics	Estonia
8	2006 Winter Olympics	Estonia at the 2004 Summer Olympics
9	Cross-country skiing at the 2006 Winter Olympics	Estonia at the 2000 Summer Olympics
10	New Zealand at the 2006 Winter Olympics	Soviet Union at the Olympics

达文档的主题，和人工标注的概念更加接近。ESA 和 CRM 模型都取得了不错的效果，CRM 在大部分指标上要优于 ESA 模型。

4 总结

维基百科涵盖了大部分日常使用的概念，是建立通用概念库的绝佳资源。通过对维基百科文档的筛选可以得到一个大规模的通用概念知识库。本文提出了一种基于排序学习的方法来实现文档的维基百科概念自动标注。首先人工对一定规模的文档进行概念标注建立训练集合，利用 Ranking SVM 排序学习算法得到对概念排序的模型，利用这个概念的排序模型对任意的文档进行概念标注。然后，利用机器学习方法自动从文档中发现概念。实验表明，本文中的方法比传统的文档概念映射方法在各类指标上都有相当大的提高，这个标注系统比传统的方法得到的概念标注更加接近人类的概念标注。

利用本文方法建立的文本的维基百科概念标注，可以运用到文本间相关度计算、文本分类、文本聚类和信息检索等自然语言处理任务中，在一定程度上克服“词袋”模型的不足，进一步提高计算机的自然语言理解能力。

参考文献

[1] Loh S, Wives L K. Concept-based text mining. Handbook of research on text and Web mining technologies, Information Science Reference, 2009

[2] Mihalcea R, Csomai A. Wikify!: linking documents to encyclopedic knowledge // Proceedings of the sixteenth ACM conference on Conference on information and knowledge management (CIKM’07). Lisbon, Portugal, 2007: 233–242

[3] Milne D, Witten I H. Learning to link with Wikipedia // Proceedings of the 17th ACM conference on Information and knowledge management (CIKM’08). Napa Valley, CA, 2008: 509–518

[4] Maron M E. On indexing, retrieval and the meaning of about. Journal of the American Society for Information Science, 1977, 28(1): 38–43

[5] Medelyan O, Witten I H, Milne D. Topic indexing with Wikipedia // Proceedings of the AAAI 2008 Workshop on Wikipedia and Artificial Intelligence (WIKIAI’08). Chicago, 2008: 19–24

[6] Ferragina P, Scaiella U. TAGME: on-the-fly annotation of short text fragments (by wikipedia entities) // Proceedings of the 19th ACM international conference on Information and knowledge management (CIKM’10). Toron to, 2010: 1625–1628

[7] Kulkarni S, Singh A, Ramakrishnan G, et al. Collective annotation of Wikipedia entities in web text // Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD’09). Paris, 2009: 457–466

[8] Gabrilovich E, Markovitch S. Wikipedia-based semantic interpretation for natural language processing. Journal of Artificial Intelligence Research, 2009, 34(1): 443–498

[9] Li Hang. Learning to rank for Information retrieval and natural language processing. xx: Morgan & Claypool Publishers, 2011

[10] Cao Yunbo, Xu Jun, Liu Tieyan, et al. Adapting ranking SVM to document retrieval // Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR’06). Seattle, 2006: 186–193