一种基于稀疏编码的语义标注方法

陈叶旺 李海波 余金山 陈维斌
(华侨大学计算机科学学院 厦门 361021)

摘 要 语义标注是实现语义网的一个重要研究内容,目前已有很多标注方法取得了不错的效果。但这些方法几乎都没有注意到本体所描述的知识往往稀疏地分布在文档中,也未能有效地利用文档的组织结构信息,使得这些方法对质量较差的文档的标注不理想。为此提出了一种基于稀疏编码的本体语义自动标注方法(Semantic Annotation Method based on Sparse Coding,SAMSC),该方法先按本体知识描述从文档中识别出一定的语义作为初始值,再通过迭代解析文档段落结构和描述主题,完成本体知识与文档资源的相关系数矩阵计算,最后在全局文档空间中通过最小化损失函数来实现用本体对文档的语义标注。实验表明,该方法能有效地对互联网中大量良莠不齐的文档进行自动语义标注,对质量差的文档资源能取得让人接受的结果。

关键词 本体,语义标注,段落结构,SAMSC
中图法分类号 TP301 **文献标识码** A

Semantic Annotation Method Based on Sparse Coding

CHEN Ye-wang LI Hai-bo YU Jin-shan CHEN Wei-bin
(College of Computer Science and Technology, Huaqiao University, Xiamen 361021, China)

Abstract Semantic annotation plays a significant role in Semantic Web research. There are many annotation methods for unstructured documents today. However, none of them takes notice of the fact that the knowledge locates in documents sparsely, and few of them make use of the structure of a document effectively, which results in that they cannot annotate document well in case that the quality of the document is poor. In this paper, we proposed a Semantic Annotation Method based on Sparse Coding(SAMSC) for unstructured data. This method starts from initiation by identifying some semantics described in documents by ontology; secondly, in order to determine the correlation between a document and a semantic topic described in ontology, it resolves the paragraph structure and topics of the document iteratively; finally, this method annotates the documents in the global range of all documents by minimizing loss function. The experiment results demonstrate the performance of this method annotates unstructured documents well in the Web automatically and effectively. Also, it annotates low quality documents better than other methods.

Keywords Ontology, Semantic annotation, Text paragraph structure, SAMSC

1 引言

语义网[1]以及本体技术的提出,使得基于本体的语义检索得到了广泛的关注。语义网技术是通过本体规范地表达领域知识,使得计算机可识别和处理。为文档提供语义标注[2,4]是实现语义网的重要一步。目前一般使用人工方式对小规模文档资源进行标注[2-5],但是面对海量数据,这种方式显得无能为力。因此,自动或半自动工具相继出现[6,7],这些工具主要应用传统信息抽取技术的方法、基于本体的信息抽取方法、基于自然语言处理的方法等,取得了不错的效果。

然而,目前现有的大量网络资源以非结构化的文档形式存在,如网页、博客、论坛帖子等。这些文档是为人类阅读准备的,而不是为了机器处理。这些文档按人类的逻辑思维结构描述特定的领域知识,其逻辑结构体现在文本的段落分布中。文档所描述的知识关系也往往稀疏地分布在不同的段落或段落的不同位置。另一方面,网络中的这些文档质量良莠不齐,差异较大。有些质量较好的文章由专业组织或专家撰写,容易通过计算机识别和进行语义标注工作。

而也有类似存在于论坛中的各种业余的探讨性帖子,句法使用不规范,文章结构随意,内容描述也不一定准确。对于这种质量较差的文本,很难自动识别和标注。

现有的语义标注方法虽然有很多,但总的来说,这些方法几乎都没有注意到本体所描述的知识往往稀疏地分布在文档中,也未能有效地利用文档的组织结构信息,使得这些方法对质量较差的文档标注结果不理想。再者,这些方法有的不注重而有的过分依赖本体的知识关系结构,标注结果受本体质

到稿日期:2010-10-11 返修日期:2011-01-28 本文受福建农业科技重大项目(2010N5008),福建省自然科学基金(A0810013)资助。
陈叶旺 博士,讲师,主要研究方向为语义检索、数据挖掘,E-mail:ywchen@hqu.edu.cn;李海波 博士,副教授,主要研究方向为企业信息化、软构件、工作流;余金山 教授,主要研究方向为软件工程、人工智能;陈维斌 教授,主要研究方向为数据库技术、数据仓库与决策支持、面向对象技术及其应用。

量影响很大。

针对这些问题，本文提出了一种基于稀疏编码的本体语义自动标注方法（Semantic Annotation Method based on Sparse Coding，SAMSC）。考虑到本体语义知识关系在文档描述中的稀疏性，先按本体知识描述从文档中识别出一定的语义作为初始值，再通过 SAMSC 模型迭代解析文档结构和描述主题，完成本体知识关系与文档资源的相关系数，最后在全局文档空间中通过最小化损失函数来实现用本体对文档的语义标注。该方法能有效地对互联网中大量以网页等形式存在的、质量良莠不齐的多种类文档知识资源进行有效的自动语义标注。实验表明，该方法不仅对于质量较好的文档标注结果好，对于质量差的文档资源也能取得让人接受的结果，即该方法对文档质量的抗干扰能力较强。

本文第 2 节是语义标注的相关工作；第 3 节介绍语义标注概念；第 4 节介绍基于 SAMSC 模型的语义标注方法；第 5 节是实验与评测；最后是总结。

2 国内外相关工作

目前已经有了一些关于语义标注的研究工作，包括应用传统信息抽取技术的方法（Information Extraction，IE）、基于本体的信息抽取方法（Ontology Based Information Extraction，OBIE）、基于自然语言处理的方法（Natural Language Processing，NLP）等。现将这些方法简单介绍如下。

1）应用传统信息抽取技术的方法。一些研究者应用传统信息抽取技术对网页进行语义标注。例如 Amileare[6] 应用机器学习的方法在标注好的训练集上进行训练，通过提供不同领域的标注训练文档，可以适应多个领域。S-Cream[3] 综合运用了 Ont-O-Mat[7] 与 Amilcare[6] 工具，将 Amilcare 产生的标注结果与 Ont-O-Mat 定义的关系元数据内的概念标记匹配。但是由于二者之间表示的差异，这种匹配十分困难，因此难以产生符合领域概念的标注结果。

2）基于本体的信息抽取方法。OBIE 方法将本体作为信息抽取过程中可用资源的一部分，比如利用本体内已有的实例信息来构造列表，简化抽取过程中对于概念实例的识别。例如 SemTag[12] 首先在 TAP 本体实例集合中查找所有与待标注词匹配的可能实例集合，然后根据待标注词的上下文与实例集合中每个实例的上下文，分别构造各自的文本向量，进行相似度计算，找到与待标注词最匹配的实例。而 KIM[5] 则可以在应用现有本体实例集对文本进行标注的同时，进行新实例的生成工作。OBIE 方法的共性问题关注的是对文本中实例的标注与提取，而未能很好地抽取出实例间的关系。

3）有的研究利用语义相似性来代替相关性。利用语义相似性进行扩展的目的，是尽可能同时提高标注的准确度，即把用来标注的本体实例扩展到与它的语义相关度大的其它实例。相似度大的概念往往相关度也很大，从而也有助于提高准确度。计算相似性的方法有很多，如余弦相似度、Dice 相似度等，这些方法的前提假设是词语之间是完全独立的。相关度计算的主要途径是利用文档集中词语间共现性的统计数据。这种方法来源于一种直觉，即在文档中经常共同出现的词语往往相关度很大。分析共现性时，可以采用词语粒度、短语粒度[14]、概念粒度[15,9] 等。另外，信息熵[16]、句法上下文[8,17] 等也是相关度计算的依据。然而，这些方法大都是利用文档或文档片断中包含的内容信息，忽略了从文档外部观察文档、实例、实例上下文之间的关系。

4）国内也有较多的关于语义标注的研究。类似于 DOM[13]，文献[10,11]提出基于分析页面结构信息的方法。文献[11]基于结果模式的 Deep Web 数据抽取机制，将数据抽取工作分为结果模式生成和数据抽取两个阶段，属性语义标注放在结果模式生成阶段来完成，有效解决了重复语义标注问题；同时针对嵌套属性问题，提出了一种有效的解决方法。

3 语义标注

简单来说，语义标注就是在本体中的知识点和资源之间建立关联，表示这些资源是这个知识点的一个扩展描述。资源可以是文本，可以是图片，也可以是多媒体的影像资料等。经过本体知识标注后的文档和知识构成一种关系，这种关系从另一个角度说明了文档与本体知识的相关程度。

利用本体可以为领域知识定义类型、描述关系，建立一个知识库及其知识管理基础设施，为传统信息抽取技术提供格式参考，同时解决无类型问题。这为上述困难提供了解决方案。本文利用本体对资源的语义标注方法有 3 个基本要素，包括标注对象、标注知识和标注方式。

（1）标注对象：指各种信息资源，包括各类有格式和无格式的文本文档，即资源库（DS）。

本文中的标注对象主要指文本文档。资源库（DS）记为 $DS=\{d_1,\cdots,d_i\cdots,,d_m\}$，其中 d_i 是第 i 个文档，m 为文档数量，$0<i<m$。

（2）标注知识：指描述在一个或多个领域本体中的概念、实例或关系。这些概念、实例或关系稀疏地分布在文档库文档中，是这些文档的描述内容。

文献[20]认为语义标注是有针对性地依赖于某个特定的知识，即领域知识，而非普适知识。实例则表述的是具体的某项领域知识或关系的一个真实存在，是我们用来标注文档的知识源。

（3）标注方式：指按标注保存形式划分。目前的标注方式主要有两种：第一种是内嵌式的，这种方式比较容易实现。但有两个不足之处，（a）内嵌标注方式已经把标注作为本体内容的一部分，修改起来比较困难，不适合用来实现用户自定义的、动态的语义标注。（b）内嵌标注方式把本体变得复杂，也增加了本体知识库维护的负担。因而，本文的标注方法采用非内嵌方式来标注文档，标注结果存储在数据库中，而不是创建一个标注本体。

4 语义标注方法

4.1 基本方法

本文的标注方法主要是针对文本内容，所以对于非纯文本格式文件，我们有针对不同类型文档的解析工具，用来提取出文档文本内容。一些系统使用简单的标注方法，只通过单个本体实例的标签值在文档文本内容中出现的次数来标注文档。这种方法无法体现词汇在文本中的作用程度，所以逐渐被更精确的词频代替。然而，这些方法还是有许多不足：（1）有些词汇虽然是指同一个意思，却常有不同的表达词组，如果只用标签值来做统计，往往会造成统计结果不准确；（2）割离

了本体知识的存在语义环境，甚至可能会使标注的结果完全错误；(3)忽略了文档本身的结构。

4.2 基于稀疏编码的语义标注(Semantic Annotation Method based on Sparse Coding, SAMSC)

现有的大量网络资源以非结构化的文档形式存在，如网页、博客、论坛帖子等，而且呈爆炸式地增长。这些文档是为人类阅读准备的，而不是为了机器处理。这些文档按人类的逻辑思维结构描述特定的领域知识，其逻辑结构体现在文本的段落分布中。文档所描述的知识关系也往往稀疏地分布在文档的不同段落或段落的不同位置。另一方面，网络中的这些文档质量良莠不齐，有些由专业组织或专家撰写的技术文章，质量好，容易通过计算机识别和进行语义标注工作。而也有类似存在于论坛中的各种业余的探讨性帖子，句法使用不规范，文章结构随意，内容描述也不一定准确。对于这种质量较差的文本，很难自动识别和标注。面对这些稀疏地描述知识且质量良莠不齐的海量信息资源，为能有效地进行标注，本小节提出一个基于 SAMSC 模型的语义标注方法。从分析文档段落结构和主题语义两方面入手，通过迭代方式求得本体知识项与文档之间的相关系数近似解，最后在全局文档空间中通过最小化损失函数来实现语义标注。

4.2.1 SAMSC 模型

在介绍 SAMSC 模型方法之前，先介绍一些基本概念。

定义 1(语义主题，Topic) 一个语义主题(Topic) t 是指某个实例 e 在领域本体知识 O 中的关系集合，以本体三元组表示为

$$t(e)=\{triple \mid (triple.subject=e) \vee (triple.object=e)\}$$

即说明语义主题可以表示为本体三元组空间的一个分布。那么一个领域本体描述了一个主题空间，记为 $E=\{\vec{t_1},\cdots,\vec{t_n}\}$。

通常，一篇文本资源由若干段落构成，每个段落又由若干字符组成。令 $\Im$ 为一个有限字符集，对于中文来说，它是指标准中文编码库，如 GB2312-80；对于英文来说，它指的是 ASCII 码或 UNICODE 所对应的字符。令 ζ 为自然语句段落分隔符集，且 $\varsigma \in \Im$，则文本段落可做如下定义。

定义 2(段落，PG) 段落 PG 为一个有限序列字符集，其结尾字符为段落分隔符，即 $PG.end \in \varsigma$。表达若干个相关的语义主题，表示为

$$PG \cong \sum_{i=1}^{n} \theta_i^{PG} \cdot \vec{t}_i$$

式中，n 为领域本体所描述的主题空间中的主题数量；$0 \leqslant \theta_i^{PG} \leqslant 1$ 表示段落 PG 与第 i 个主题之间的相关度系数；$\vec{t}_i \in E$。通常，一个段落所要表达的内容比较有限，因而可以认为 $\vec{\theta^{PG}}=\{\theta_1^{PG},\cdots,\theta_n^{PG}\}^{\mathrm{T}}$ 是一个稀疏矩阵。

定义 3(文本资源，Text) 文本资源 Text 为一个有限序列段落集。

为评判一篇文本资源与领域本体所描述的主题空间 E 的关系，定义损失函数：

$$lf=\| P-E\Theta \|_F^2 \tag{1}$$

式中，P 为这篇文档资源所包含的 k 个有序段落集合，即 $P=\{PG_1,\cdots,PG_k\}$；Θ 为文本中各段落与语义主题的相关系数矩阵，即

$$\Theta=\{\vec{\theta^{PG}}_1,\cdots\cdots,\vec{\theta^{PG}}_k\}$$

在一篇文本中，某些段落与段落之间暗含有上下承接关系，描述的主题相互紧密关联或相近，这些体现为一篇文章的结构特征。为捕捉第 i 个段落 PG 与其前述段落之间的这一特征，定义结构特征函数：

$$struc(PG_i)=\| \vec{\theta^{PG_i}}-\sum_{j=1}^{j=i-1}\gamma_j^{PG_i}\cdot\vec{\theta^{PG_j}} \|_F^2 \tag{2}$$

式中，$\gamma_j^{PG_i}$ ($\gamma_j^{PG_i} \in [0,1)$)为第 i 个段落与第 j 个段落之间相关的系数，表示第 j 个段落对第 i 个段落的影响程度。一篇有 k 个段落的文章，段落之间的相互关系可用一个 $k \times k$ 三角矩阵的 γ 值表示。

根据以上介绍，若给定一篇包含有 k 个段落的资源文档，其段落的有序集合为 P 和一个本体知识库所描述的主题空间集合 E。我们给出一个文档资源的语义标注模型 SAMSC (Semantic Annotation Model for Text)：

$$\begin{aligned} samt &= lf+\lambda\times\sum_{i=1}struc(PG_i) \\ &= \| P-E\Theta \|_F^2+\lambda\times\sum_{i=1}\| \vec{\theta^{PG_i}}-\sum_{j=0}^{j=i-1}\gamma_i^{PG_j}\cdot\vec{\theta^{PG_j}} \|_F^2 \\ &= \| P-E\Theta \|_F^2+\lambda\times\sum_{i=1}\| \vec{\theta^{PG_i}}-\vec{\gamma_i}\cdot\sum_{j=0}^{j=i-1}\vec{\theta^{PG_j}} \|_F^2 \end{aligned} \tag{3}$$

式中，$\vec{\gamma_i}$ 表示 γ 矩阵中的第 i 列；多项式中的第一项意义在于能够重新调整各段落与本体知识主题之间的关系；第二项则是用于估计段落之间的相互关系。这样，既能够捕获一篇文档资源的语义信息，又可以有效计算文档的结构关系。

假设在 M 个文档的数据集中，共享主题矩阵 X 通过最小化以下的损失函数以得到全局的最优解。

$$\min_{X,\{\Theta^{(n)}\},\{\gamma^{PG_i(n)}\}}\sum_{n=0}^{M}samt^{(n)}$$

4.2.2 SAMSC 模型求解

本小节介绍 SAMSC 模型的近似解法。为了方便计算，首先把目标函数写成矩阵形式。通过最小化这个模型取值来估算文档与主题空间 E 的系数矩阵 Θ 以及段落结构关系。

当 E 与 Θ 确定时，对 $\vec{\gamma_i}$ 求偏导，有

$$\frac{\partial(samt)}{\partial(\vec{\gamma_i})}=-\lambda\cdot(\vec{\theta^{PG_i}}-\vec{\gamma_i}\cdot\sum_{j=0}^{j=i-1}\vec{\theta^{PG_j}})\times\sum_{j=0}^{j=i-1}(\vec{\theta^{PG_j}})^{\mathrm{T}} \tag{4}$$

令 $\frac{\partial(samt)}{\partial(\vec{\gamma_i})}=0$，则当 $i>0$ 时，有

$$\vec{\gamma_i}=\frac{\vec{\theta^{PG_i}}}{\sum_{j=0}^{j=i-1}\vec{\theta^{PG_j}}} \tag{5}$$

当 $\vec{\gamma^{PG_i}}$ 和 E 确定时，对 $\vec{\theta^{PG_i}}$ 求偏导，有

$$\begin{aligned} \frac{\partial(samt)}{\partial(\vec{\theta^{PG_i}})} &= -(\vec{t_i})^{\mathrm{T}}\times(P-E\Theta)+\lambda\times(\vec{\theta^{PG_i}}-\vec{\gamma_i}\cdot\sum_{j=0}^{j=i-1}\vec{\theta^{PG_j}}- \\ &\quad \sum_{j=i+1}\vec{\gamma_j}(\vec{\theta^{PG_j}}-\vec{\gamma_j}\cdot\sum_{m=0}^{m=j-1}\vec{\theta^{PG_m}})) \end{aligned} \tag{6}$$

又因为 $E\Theta=\vec{t_i}\cdot\vec{\theta^{PG_i}}+\sum_{j=0}^{i-1}\vec{t_j}\cdot\vec{\theta^{PG_j}}+\sum_{j=i+1}\vec{t_j}\cdot\vec{\theta^{PG_j}}$，所以上式可改写为

$$\begin{aligned} \frac{\partial(samt)}{\partial(\vec{\theta^{PG_i}})} &= -(\vec{t_i})^{\mathrm{T}}\times(P-\sum_{j=0}^{i-1}\vec{t_j}\cdot\vec{\theta^{PG_j}}-\sum_{j=i+1}\vec{t_j}\cdot\vec{\theta^{PG_j}})-\lambda\times\vec{\gamma_i} \\ &\quad \cdot\sum_{j=0}^{j=i-1}\vec{\theta^{PG_j}}-\lambda\times\sum_{j=i+1}\vec{\gamma_j}(\vec{\theta^{PG_j}}-\vec{\gamma_j}\times(\sum_{m=0}^{m=i-1}\vec{\theta^{PG_m}}- \\ &\quad \sum_{m=i+1}^{m=j-1}\vec{\theta^{PG_m}}))+\vec{\theta^{PG_i}}\times(\lambda+(\vec{t_i})^{\mathrm{T}}\times\vec{t_i}+\lambda\times\sum_{j=i+1} \\ &\quad (\vec{\gamma_j})^2) \end{aligned} \tag{7}$$

当令 $\frac{\partial(samt)}{\partial(\vec{\theta^{PG_i}})}=0$ 时，可得

$$\overrightarrow{\theta^{PG_i}}=\frac{-G}{\lambda+(\vec{t}_i)^T\times\vec{t}_i+\lambda\times\sum_{j=i+1}(\vec{\gamma}_j)^2} \tag{8}$$

其中，

$$G=-(\vec{t}_i)^T\times(P-\sum_{j=1}^{i-1}\vec{t}_j\cdot\overrightarrow{\theta^{PG_j}}-\sum_{j=i+1}\vec{t}_j\cdot\overrightarrow{\theta^{PG_j}})-\lambda\times\vec{\gamma}_i\cdot\sum_{j=0}^{j=i-1}\overrightarrow{\theta^{PG_j}}-\lambda\times\vec{\gamma}_j\sum_{j=i+1}(\overrightarrow{\theta^{PG_j}}-\vec{\gamma}_j\times(\sum_{m=0}^{m=i-1}\overrightarrow{\theta^{PG_m}}-\sum_{m=i+1}^{m=j-1}\overrightarrow{\theta^{PG_m}}))$$

当确定 Θ 和所有的 $\overrightarrow{\theta^{PG}}$ 时，假设 D 是所有 M 篇文档中所描述的领域本体三元组矩阵，我们可以通过以下公式优化 E。

$$E=D_M\Theta^{-1} \tag{9}$$

SAMSC 的近似求解过程：首先依据本体知识给 E 赋初值，正规化矩阵 E；再对 Θ 做初始化，其初始化过程在下一小节中介绍。γ 的初始值全设为 0，表示假设段落之间完全独立。然后重复以上优化过程 c 轮。在每一轮中，先优化所有线程中的 γ，再利用式(5)、式(8)优化所有线程中的 Θ，利用式(9)顺序地优化 E，并将其正规化。算法如 Algorithm 1 所示。

Algorithm 1 SAMSC 解法

```
Input: D for m Documents, Θ
Output: optimized Θ, γ, E
1. Initialization: Randomize E;
2. Normalize E;
3. repeat
4.   for each document do
5.     repeat
6.       optimize γ by equation(7);
7.       if γ>1 then
8.           γ=1;
9.         break;
10.        end;
11.      until converged;
12.      while not converged do
13.        optimize Θ by equation(10);
14.        if θ>1 then
15.          θ=1;
16.        break;
17.      end;
18.    end
19.  end
20.  repeat
21.    optimize E by equation(11);
22.  until converged;
23. until c round;
```

复杂度分析：文档集输入规模为 m；设 Θ 输入规模为 n，即有 n 个语义主题。算法 Algorithm1 的基本操作是步骤 6、13 和 21。设一篇文档段落数为 k，则步骤 6 和 9 的复杂度为 $O(k)$；对于步骤 21 来说，复杂度取决于 Θ 的逆矩阵，通过 LU 分解求逆的复杂度为 $O(n^2)$。因而总体上来说复杂度为 $O(m*(n^2+k))$，即最后复杂度为 $O(m*n^2)$。

4.2.3 语义主题与文档相关度

由于最终标注结果要表示成文档 d 与语义主题 t 之间的关系，而非段落与语义主题之间的关系，因此需要将 Θ 中的每行矢量做综合。Θ 中第 i 行的值表示一篇文档的各个段落与第 i 个语义主题的相关性，即 $\vec{\theta}_i=(\overrightarrow{\theta_i^{PG_1}},\overrightarrow{\theta_i^{PG_2}},\cdots,\overrightarrow{\theta_i^{PG_k}})$。若假设 $\vec{\theta}_i$ 是离散的概率，那么依据贝叶斯公式有

$$P(t_i|d)=\frac{P(t_i)\times\sum_{j=1}^{k}P(PG_j|t_i)}{P(d)}=\frac{P(t_i)\times\sum_{j=1}^{k}(\overrightarrow{\theta_i^{PG_j}})}{P(d)} \tag{10}$$

式中，$P(t_i|d)$ 表示文档 d 与主题 t_i 之间的相关概率；k 是文档 d 的段落数；$P(d)$ 表示文档 d 出现的概率，其取值设为 $1/M$，M 为文档空间中的文档总数；$P(t_i)$ 表示选择主题 t_i 进行标注工作的概率，其取值设为 $1/n$，n 为语义主题空间中主题总数。本文方法最终使用 $P(t_i|d)$ 来表达文档 d 与主题语义主题 t_i 之间的相关度。

5 实验与评测

我们在 Java 环境下，使用 Eclipse 作为开发平台。实验运行平台配置为 Intel Centrino Duo T2400 1.83GHz PC、2GB 内存、WindowsXP SP2。

测试数据：为了考察我们方法的有效性和正确性，实验中建立 3 个不同规模的本体。第一个是农作物病虫害领域本体 CropDisease，第二个是花卉知识本体 Flower，第三个是足球本体 Soccer。统计信息见表 1。

表 1 本体统计数据

领域本体	本体概念总数	实例总数
Soccer	54	1,230
Flower	113	2,400
CropDisease	274	3,730

对应于 3 种不同的本体，采用的检索文档集有 3 个：新浪足球新闻网国际足球新闻；花卉知识文档，来自上海花卉网[1)]；来自中国农科院作物品种资源研究所依据《中国粮食作物、经济作物、药用植物病虫原色图鉴》、《中国农业百科全书》所制作的农作物病虫害知识[2)]。其中，新浪足球新闻属于即时新闻报道，文章创作时间较短，不可能长时间地检验与校对。因而相对于后两者来说，其文章的严谨性和质量相对较差。花卉知识文档次之，中科院制作的病虫害知识经过多年实践检验，且文字表达科学严谨，质量最好。

我们将这些文档知识分别按不同格式(txt、html、xml、doc、pdf)转化为纯文本文档。转化后的文档集的统计信息见表 2。

表 2 文档集的统计数据

领域文档集	文档数	段落总数	句子总数	每篇平均词组数
花卉知识	619	4,233	17,523	407
新浪国际足球新闻	703	5,013	36,351	523
农作物病虫害知识	1,119	5,452	70,743	495

评测标准和结果分析：在测试数据集进行语义标注，每个有标注结果的实例都有相对应的被标注文档，我们对这些文档进行人工统计、分析。以人工处理结果作为比较基准，对方法的有效性进行评价。另外，我们将其与 Deep Web[10,11] 方法做比较。

最终经过优化之后，用于标注文档的实例统计结果如表 3 所列，而文档标注统计结果如表 4 所列。由表 3 可知，并非所有本体实例都会参与到标注过程，中间有一些实例会被过

1) http://www.flower-sh.cn/article_list.asp? c_id=74&page=4

2) http://icgr.caas.net.cn/disease/

滤，主要是匿名或以对象属性存在的实例。

表 3　最终用于标注的实例数据统计

领域本体	实例总数	最终用于标注的实例总数
Soccer	1,230	1,030
Flower	2,400	2,032
CropDisease	3,730	3,314

表 4　标注结果统计

领域文档集	平均一个文档被标注的实例数量		平均一个实例标注的文档数量	
	Deep Web	SAMSC	Deep Web	SAMSC
花卉知识	3.6	2.9	0.18	0.15
新浪国际足球新闻	5.4	4.6	1.76	1.47
农作物病虫害知识	3.0	2.4	0.32	0.3

实验中有多个本体，为进行有效评测，在测试时，对于不同本体我们选用与其对应的文档集进行标注。根据文档规模，要完全从人力上判断是难以完成的。因此，采用 3 个评判标准：

(1)任意取 $n(n=20)$ 个有标注结果的本体实例的前 $k(k=10)$ 个标注结果，采用 $Precision[ND\text{-}DOC]@(n,k)$ 来衡量前 k 个被标注文档的准确率，按式(11)计算。

$$Precision[\text{IND-DOC}]@(n,k)=\frac{1}{n}\times\sum_{i=0}^{i=n}(\#\text{ of relevant docs in top-}k\text{ annotated of }e_i/k) \quad (11)$$

(2)任意取 $n(n=20)$ 个被标注过的文档，采用 $Recall[\text{DOC-IND}]@\{n,t\}$ 来衡量与文档相关度大于 $t(t=0.40)$ 的本体实例的查全率，按式(12)计算。

(3)任意取被标注过的文档 $n(n=20)$ 个和有标注结果的本体实例各 $m(m=20)$ 个，采用 $F\text{-}Measure@(m,n,k,t)$ 来衡量均衡性，其中 k 为前 k 个被标注文档的准确率，$t=0.4$，按式(13)计算。

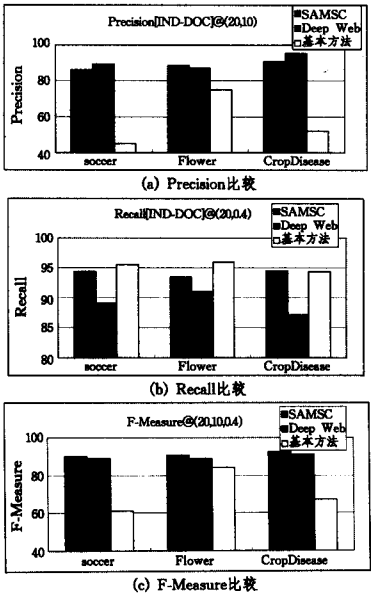
$$Recall[\text{DOC-IND}]@\{n,t\}=\frac{1}{n}\times\sum_{i=0}^{i=n}\frac{\#\text{ of individuals whose revelant degree to doc}_i\text{ are bigger than }t\text{ got by our method}}{\#\text{ of individuals whose revelant degree to doc}_i\text{ are bigger than }t\text{ got mannually}} \quad (12)$$

$$F\text{-}Measure@(m,n,k,t)=\frac{2\times(Precision[\text{IND}-\text{DOC}]@(m,k)\times Recall[\text{DOC}-\text{IND}]@(n,t))}{Precision[\text{IND}-\text{DOC}]@(m,k)+Recall[\text{DOC}-\text{IND}]@(n,t)} \quad (13)$$

这样，$Recall[\text{DOC-IND}]@(n,t)$、$Precision[\text{IND-DOC}]@(m,k)$ 和 $F\text{-Measure}(m,n,k,t)$ 一起能够更全面地对 top-n 的检索结果进行评价，而这也符合大多数检索用户的习惯，因为大多数用户在检索过程中主要关注 top-n 检索结果。

5.1　实验 1

首先，用来做实验的文档是描述详细、结构清晰的，领域本体也是描述准确且充分的。实验结果与 4.1 节描述的基本方法及 Deep Web 做比较。图 1 为 $n=20,k=10,t=0.4$ 时 $Recall[\text{DOC-IND}]@(n,t)$、$Precision[\text{IND-DOC}]@(n,t)$ 和 $F\text{-}Measure(m,n,k,t)$ 的实验结果。结果表明，Deep Web 方法与文献[10]中报告结果相近；基于 SAMSC 的方法 $Precision[\text{IND-DOC}]@(n,t)$ 与 Deep Web 方法相比不相上下，而 $Recall[\text{DOC-IND}]@(n,t)$ 比 Deep Web 方法要高，接近基本方法。基本方法查准率比较低，但查全率较高。从 F-Measure 比较来看，在文档质量较好的情况下，SAMSC 方法均衡性能与 Deep Web 基本相当。



(a) Precision比较

(b) Recall比较

(c) F-Measure比较

图 1　3 种方法比较

从本项实验来看，在文档质量和本体质量都较好的情况下，SAMSC 模型与 Deep Web 表现相差不大。

5.2　实验 2

为检查文档质量对方法的影响，我们从花卉知识文档集中人工选出一些质量较好的文档，同时选出一些质量较差的文档，有的是人为地把某些质量较好的文档故意删除一部分内容，或打乱其段落和句子顺序，使其变得难以理解。实验结果如图 2 所示。从图中可明显看出文档质量对标注结果有很大影响。文档结构清晰、描述准确，标注的结果就较好，反之较差。其中，对于 SAMSC 模型来说，文档质量下降致使查准率下降约 44%，查全率下降约 28%，F-Measure 下降约 36%。然而对于 Deep Web 方法来说，文档质量下降是灾难性的，致使查准率下降约 77%，查全率下降约 69%，F-Measure 下降约 74%。

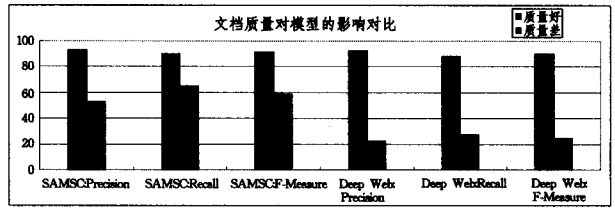


图 2　文档质量对模型影响对比图

从本项实验可以看出，当文档质量变化较大时，SAMSC 模型受影响程度要远小于 Deep Web，也就是说 SAMSC 模型对文档自身质量的抗干扰能力比 Deep Web 强。

结束语　语义网技术是通过本体规范地表达领域知识，使得计算机可识别和处理。要实现语义网的理想，实现对互联网中的各种资源尤其是文本资源做语义标注是重要的一步。本文针对网络中的文档质量差异性大、描述知识具有稀疏性的特点，提出了一种基于稀疏编码的本体语义自动标注方法(Semantic Annotation Method based on Sparse Coding, SAMSC)。该方法先按本体知识描述从文档中识别出一定的语义作为初始值，再通过迭代解析文档整体段落结构和描述主题，从而完成本体知识关系与文档资源的相关系数，最后在

(下转第 181 页)

ysis and Machine Intelligence, 2005, 27(6): 1866-1881

[10] Ayad H G, Kamel M S. Cumulative Voting Consensus Method for Partitions with A Variable Number of Clusters[J]. IEEE Transaction on Pattern Analysis and Machine Intelligence, 2008, 30(1): 160-173

[11] Avogadri R, Valentini G. Fuzzy ensemble clustering based on random projections for DNA microarray data analysis[J]. Artificial Intelligence in Medicine, 2009, 45(2/3): 173-183

[12] Ester M, Kriegel H-P, et al. A Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise[C]// The ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Portland, Oregon, 1996: 226-231

[13] Azimi J, Fern X. Adaptive Cluster Ensemble Selection[C]// Proceedings of International Joint Conference on Artificial Intelligence(IJCAI). California, 2009: 992-997

[14] Fern X Z, Brodley C E. Cluster Ensembles for High Dimensional Clustering: An Empirical Study[EB/OL]. http://Web. engr. oregonstate. edu/~xfern/clustensem. pdf, 2010-02-20

[15] 陈华辉，施伯乐. 基于随机投影的并行数据流聚类方法[J]. 模式识别与人工智能，2009，22：113-122

(上接第 154 页)

全局文档空间中通过最小化损失函数来实现用本体对文档的语义标注。基于这个模型，本文实现了一个针对非结构化文档的语义标注工具。通过这个工具所做的实验表明，该方法能有效地对互联网中大量以网页等形式存在的、质量良莠不齐的多种类文档知识资源进行有效的自动语义标注。这不仅对于质量较好的文档标注结果好，对于质量差的文档资源也能取得让人接受的结果，即该方法对文档质量的抗干扰能力较强。另一方面，本体质量的优劣同样对标注结果有重要影响。实验也同样表明，该方法受本体质量变化的影响相对较小。

通过本文的工作我们认识到，网络中文本资源质量差异很大，其中质量不好的文本资源占相当的比重。因而，我们认为，如果不能对这些质量相对较差的资源进行有效的标注，会造成网络资源丢失或浪费。另一方面，文档所描述的知识关系往往稀疏地分布在文档的不同段落中，且段落与段落之间本身有一定的关系。段落之间关系密切，则其描述的主题相同或相近的概率较大。本文的方法对知识分布稀疏的段落处理能力较好，但对关系密切段落之间还没有相应的对策，这将是我们下一步需要研究的工作。

参 考 文 献

[1] Uschold M. Converting an Informal Ontology into Ontolingua [C]// Proceedings of the Workshop on Ontological Engineering held in conjunction with ECAI 96. Budapest, March 1996

[2] Dill S, Eiron N, Gibson D, et al. A Case for Automated Large Scale Semantic Annotation[J]. Journal of Web Sematics, 2003 (1): 115-132

[3] Handschuh S, Staab S, Ciravegna F. S-cream Semi-automatic Creation of Meta-data[C]// Gómez-Pérez A, Benjamins V R, eds. 13th Intl. Conf. on Knowledge Engineering and Knowledge Management Ontologies and the Semantic Web. LNCS, Vol. 2473. Berlin Heidelberg New York: Springer Verlag, 2002: 358-372

[4] Kiryakov A, Popov B, Terziev I, et al. Semantic Annotation, Indexing, and Retrieval[J]. Journal of Web Sematics 2, 2004(1): 47-49

[5] Popov B, Kiryakov A, Ognyanoff D, et al. KIM: A Semantic Platform for Information Extaction and Retrieval[J]. Journal of Natural Language Engineering, Cambridge University Press, 2004(3/4): 375-392

[6] Ciravegna F, Wilks Y. Designing adaptive informationextraction for the Semantic Web in amilcare, HandschuhS[C]// Staab S, ed. Annotation for the Semantic Web, Frontiers in Artificial Intelligence and Applications. Amsterdam: IOS Press, 2003: 112-127

[7] Handsehuh S, Staab S, Maedche A. CREAM: Creating relational metadata with a component-based, ontology-driven annotation framework[C]// Proc. of the 1st Int'l Conf on Knowledge Capture. New York: ACM, 2001: 76-83

[8] Gregory G. Use of syntactic context to produce term association lists for text retrieval[C]// Belkin N, Ingwersen P, Pejtersen A M, eds. Proc. of the 15th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. Copenhagen: ACM Press, 1992: 89-97

[9] Chang Y, Ounis I, Kim M. Query reformulation using automatically generated query concepts from a document space[J]. Information Processing and Management, 2006, 42: 453-468

[10] Yuan L, Li Z H, Chen S L. Ontology-based annotation for deep Web data[J]. Journal of Software, 2008, 19(2): 237-245

[11] Ma An-xiang, Zhang Bin, Gao Ke-ning, et al. Deep Web Data Extraction Based on Result Pattern[J]. Journal of Computer Research and Development, 2009 46(2): 280-288

[12] Gardent C, Parmentier Y. SemTAG: a platform for specifying tree adjoining grammars and performing TAG-based semantic construction[C]// Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions. 2007: 13-16

[13] Tenier S, Toussaint Y, Napoli A, et al. Instantiation of relations for semantic annotation[C]// Proc. of Web Intelligence 2006. Los Alam itos: IEEE Computer Society. 2006: 463-472

[14] Xu J X, Croft W B. Improving the effectiveness of information retrieval with local context analysis[J]. ACM Trans. on Information Systems, 2000, 18(1): 79-112

[15] Zhang M, Song R H, Ma S P. Document Refinternet based on semantic query expansion[J]. Chinese Journal of Computers, 2004, 27(10): 1395-1401

[16] Jang M G, Myaeng S H, Park S Y. Using mutual information to resolve query translation ambiguities and query term weighting [C]// Dale R, Chuch K. eds. Proc. of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics. College Park: Association for Computational Linguistics, 1999: 223-229

[17] Gao J F, Zhou M, Nie J Y, et al. Resolving query translation ambiguity using a decaying co-occurrence model and syntactic dependence relations[C]// Järvelin K, Chairs P, Baeza-Yates R, et al. eds. Proc. of the 25th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. Tampere: ACM Press, 2002: 183-190