基于 Citation-KNN 的语义隐含主题词自动抽取方法[1]

章成志[1,2]　刘耀[1]　王惠临[1]

1. 中国科学技术信息研究所，北京，100038

2. 南京理工大学信息管理系，南京，210094

{zhangchz, liuy, wanghl}istic.ac.cn

摘　要：现有的关键词抽取技术仅仅是抽取出现在正文中的词汇，不能够抽取语义上隐含的主题。语义隐含主题的抽取是文本挖掘技术的难点。众所周知，KNN 方法作为机器学习领域的一个经典的方法，在很多领域都有出色的表现。本文以 KNN 算法为基础，提出基于 Citation-KNN 的语义隐含主题词自动抽取方法。实验结果表明该方法在进行语义隐含主题词抽取任务上的有效性。

关键词：*关键词抽取；隐含语义主题词；Citation-KNN；文本挖掘*

1 引 言

关键词是最能反映文章主题或内容的词汇，是为了满足文献标引或检索工作的需要而从文章中萃取出的、表示全文主题内容信息条目的单词、词组或术语。在文献情报领域，关键词抽取是一项困难的任务。首先，关键词抽取是一项需要高度概括、分析和创造的活动，需要标引人员具有较高的专业知识和标引经验。其次，为了准确描述文本内容，标引人员通常会创造一些由多个词按照一定规则连接起来的组合词组。这些组合词在文本中可能很少出现，甚至不出现。另外，由于经验和知识背景不同，在标引同一篇文本时，不同的标引人员会给出不同的关键词集合。这突出反映在作者给出的关键词与专业标引人员给出的关键词通常存在很大差异。

[1] 本研究受“十一五”国家科技支撑计划重点项目(2006BAH03B02)、教育部人文社科项目(06JC870001)、南京理工大学青年科研扶持基金项目(JGQN0701)、南京理工大学科研启动基金项目(AB41123)资助。

关键词自动抽取能够解决上面的第三个问题，但对于前面两个问题所起的作用很小。这是因为目前用于自然语言处理的各种机器学习方法，还不能真正理解文本内容，只能通过大量的词或词组出现的频度、句法规则等信息进行统计和学习。但是，在许多应用领域，基于大规模数据集上的机器学习方法，特别是基于统计的机器学习，通常比完全采用人工方法效果更好、更稳定，例如文本分类、OCR 识别、词义排歧、信息检索等。

H. P. Luhn 在 20 世纪 50 年代末首先开展自动标引试验 (Luhn 1957, Luhn 1958)，而在 1963 年，美国 Chemical Abstracts 从第 58 卷起，就开始采用电子计算机编制关键词索引，提供快速检索文献资料主题的途径。纯粹的统计方法最早也最常被应用于关键词自动抽取 (Edmundson & Oswald 1959, Edmundson 1969, Chien 1997)；20 世纪 70 年代初，Lois L. Earl 开始采用句法分析等语言学方法 (Lois 1970)；70 年代中期，Salton 等将机器学习技术引入关键词自动抽取中 (Salton, Wong & Yang 1975)；20 世纪 90 年代末，Turney 将遗传算法 (Turney 1999, Turney 2000)、Frank 将 Bayes 方法引入关键词自动抽取 (Frank, Paynter & Witten, et al 1999)。近年来关键词自动抽取的研究趋于活跃，2001 年，Anjewierden 与 Kabel 提出基于本体的自动标引方法 (Anjewierden & Kabel 2001)；2003 年，Tomokiyo 与 Hurst 提出了基于语言模型的关键词提取方法 (Tomokiyo & Hurst 2003)，Hulth 利用 Bagging 算法进行了基于集成学习的关键词抽取 (Hulth 2003)；2004 年，李素建提出基于最大熵模型的关键词提取方法 (李素建, 王厚峰, 俞士汶等，2004)；2007 年，Ercan G.与 Cicekli I.提出基于词汇链的自动标引方法 (Ercan & Cicekli 2007)。

根据 Turney 的研究，人工标注的词汇，大约 65%至 90％出现在正文中 (Turney 1997)。那些不出现在正文中的关键词本文称之为“隐含主题”。隐含主题的抽取是一项非常困难的工作，现有的关键词自动抽取算法，无论是基于机器学习还是基于纯粹的统计方法，都很难抽取这部分词汇。通常的隐含主题词自动抽取方法是借助于外部资源，如叙词表、本体等资源，将语义隐含主题词自动抽取过程转换为主题词的分类过程，或将文本的关键词转换为主题词。本文尝试使用 Citation-KNN 的语义隐含主题词自动抽取算法来抽取文章的隐含主题。实验证明，这一方法是切实有效的。

2 基于 Citation-KNN 的语义隐含主题词自动抽取算法

2.1 Citation-KNN 算法描述

（1） KNN 算法概述

K 最近邻方法（KNN）是一种基于统计的懒惰学习算法，是由 Cover 和 Hart 于 1968 年提出的 (Cover & Hart 1968)。KNN 方法在很多领域都有应用，在文本自动分类领域，K 最近邻方法被证明是效果最好的方法之一 (Yang & Liu 1999)。测试样本根据最

近邻中的多数类进行分类。

$$y'_i=\text{argmax}\sum_{j=1}^{K} I(v=y_j) \quad (1)$$

其中，v 是类标号，y'_i 是一个最近邻的类标号，I(·)是指示函数，如果其参数为真，则返回“1”，否则返回“0”。

由于每个近邻对分类的影响可能不一样，可以根据测试样本与每个最近邻 x_i 的相似度对最近邻进行加权 (Tan, Steinbach & Kumar, 2006)，越相似的近邻，赋予越高的权重。如果直接以测试样本与近邻的相似度作为权重，则绝对权重公式、相对权重计算公式分别为式（2）、式（3）。

$$w_i= \text{Sim}(x'_i, x_j) \quad (2)$$

$$w_i= \text{Sim}(x_i, x_j) \Big/ \sum_{j=1}^{K} \text{Sim}(x_i, x_j) \quad (3)$$

考虑到近邻的权重后，分类决策函数为：

$$y'_i= \text{argmax}\sum_{j=1}^{K} w_i \cdot I(v=y_j) \quad (4)$$

（2） Citation-KNN 算法描述

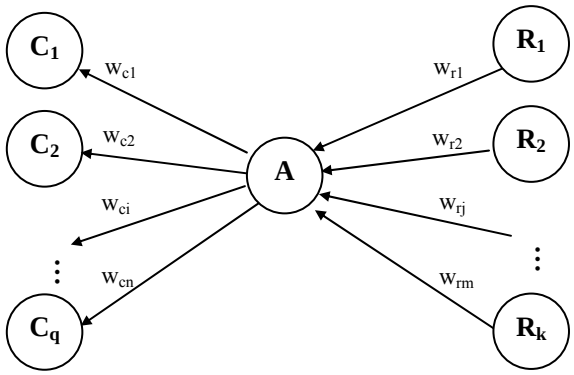

图 1　Citation-KNN 中文本“引用”与“被引用”示意图

Citation-KNN 最初由 Jun Wang 和 Jean-Daniel Zucker 提出，并用于解决多示例学习问题（Wang & Zucker 2000）。Citation-KNN 是对传统 KNN 算法的一种改进，主要思想是借助于文献计量学中的引用与被引用这一思路，在对测试样本 x'_i 进行类别决策时，除了考虑最近邻的 K 个训练样本的类别外（即测试样本的“引文”），还考虑到训练样本集中将 x'_i 作为其 K 个最近邻之一的训练样本（即测试样本的“被引”样本）

的类别。图 1 为 Citation-KNN 中文本“引用”与“被引用”示意图。

图 1 中，R_1、R_2、…、R_k 为被样本 A“引用”的 K 个文本，w_{r1}、w_{r2}、…、w_{rk} 为对应的权重；C_1、C_2、…、C_q 为样本 A 的“被引”Q 个文本，w_{C1}、w_{C2}、…、w_{Cq} 为对应的权重。

在进行分类决策时，综合考虑“引用”与“被引”的最近邻样本的类标号，并可以根据“引用”与“被引”对分类的影响赋予不同的权重，因此给出如下的分类决策函数：

$$y'_i = \text{argmax}\left(\sum_{i=1}^{K}(\alpha \cdot w_{r_i} \cdot I(v = R_i)) + \sum_{j=1}^{Q}(\beta \cdot w_{c_j} \cdot I(v = C_j))\right) \quad (5)$$

其中，α、β 分别为 Citation-KNN 中“引用”与“被引”的权重，且 α+β=1，本文取 α=β=0.5。

图 2 给出了 Citation-KNN 算法的具体描述。

算法：Citation-KNN 算法描述

输入：测试集$\{(x'_1,y'_1),\ldots,(x'_n,y'_n)\}$，,其中 $x'_i \in X'$，$y'_i \in Y$，训练集$\{(x_1,y_1),\ldots,(x_m,y_m)\}$，其中 $x_j \in X$, $y_j \in Y$

输出：测试集中每个 x'_i 对应的类标号 y'_i

步骤：

设定最近邻的数目 K，设定 α=β=0.5；

For i=1 to N

在训练集中选择离 x'_i 最近的 K 个训练样本构成的集合 $\mathbf{x_1}$，并计算每个样本的权重 w_{ri}，1<i<K；

在训练集中选择以 x'_i 作为其最近的 K 个最近邻之一的训练样本集合 $\mathbf{x_2}$；并计算每个样本的权重 w_{cj}，1<j<Q；

$$y'_i = \text{argmax}\left(\sum_{i=1}^{K}(\alpha \cdot w_{r_i} \cdot I(v = R_i)) + \sum_{j=1}^{Q}(\beta \cdot w_{c_j} \cdot I(v = C_j))\right)$$

End For

图 2　Citation-KNN 算法描述

2.2　Citation-KNN 中的近邻加权方法

本文利用 Citation-KNN 进行隐含主题词自动抽取中，根据不同近邻的特征对决策函数

进行加权。主要用到的加权方法有：根据相似度大小进行加权、根据样本本身的特征（如 PageRank 值、引用频次等）进行加权。其中相似度加权公式如式（3）所示，相似度为文本向量夹角的余弦 （Baeza-Yates & Ribeiro-Neto, 1999）。样本的 PageRank 值、引用频次的定义与计算方法见文（Zhang, Su & Zhou 2008）。

2.3 基于 Citation-KNN 的语义隐含主题词自动抽取算法

本文将隐含主题词自动抽取转化为分类学习问题，根据图 2 所示的 Citation-KNN 算法，得到待抽取文档的 K 个相似近邻的样本文档集合与将待抽取文档作为其最近的 K 个最近邻之一的训练样本文档集合，结合每个样本文档的权重，进行投票，最终得到待抽取文档的隐含主题词自动抽取结果。

3 实验结果分析与讨论

3.1 试验数据与评价方法

（1）试验数据

实验使用的数据集是中国学术期刊全文数据库[1]。从中国学术期刊全文数据库经济类数据中选出由作者给出了关键词的文献作为 K 最近邻关键词抽取的训练集，共 10 万余篇，从中国学术期刊全文数据库 2005 年数据中随机选出作者标注关键词的 600 篇文献作为测试集。

（2）评价方法

实验结果的评价采用 Turney 提出的方案，使用准确率（*Precision*）和召回率（*Recall*）以及 F_1 衡量算法的性能。在 Turney 的方案中，机器抽取的关键词和人工标注的关键词完全一致才算匹配 (Turney 1997)。定义：

$$Precision=\frac{a}{b} \quad (6)$$

$$Recall=\frac{a}{c} \quad (7)$$

$$F_1=(2\times Precision\times Recall) / (Precision + Recall) \quad (8)$$

其中，a 是机器抽取的关键词和人工标注的关键词完全匹配的数目，b 是机器自动

[1] 中国期刊全文数据库. http://www.cnki.net. Accessed: 2007.10.10.

标注的关键词数目，C 是人工标注的关键词数目。

3.2 试验结果与分析

本文进行基于 KNN 方法与基于 Citation-KNN 的隐含主题词自动抽取的对照研究。依据文本向量夹角的余弦作为文本间的相似度。表 1 为其中一篇篇名为“现代网络银行发展中的金融监管思考”的文章的相似文献集合前 10 篇最相似的文献信息。

篇名	中文关键词	引用频次	PageRank 值
网络银行发展中的问题及其对策	网络银行，金融电子化，金融监管	1	0.575000
网络银行理论机器在我国的实践	网络银行，理论依据	1	0.377679
全球网络银行的发展与中国网络银行发展战略	网络银行，生成机理，制约因素，发展战旅	1	0.510606
对我国网络银行发展与监管问题的研究	网络银行，监管	3	3.219107
网络银行的竞争有事探析	网络银行，竞争优势，阻碍因素，政策建议	1	0.510606
网络银行的安全性分析	网络银行，安全性	1	0.320000
国外网络银行发展模式的启示	网络银行，网络安全，发展模式，启示	1	0.362500
西方网络银行的发展战略及启示	网络银行，发展战旅，启示	1	0.227273
我国网络银行集约化经营之策略	网络银行，集约化经营，网上支付，便利服务，网络顾客	1	0.362500
我国发展网络银行所面临的问题与对策	网络银行，创新，对策	1	0.433333

表 1 相似文档(Top-10)样例

本文对隐含主题词自动抽取的测评方法为计算标引结果的查准率、召回率以及 F_1 值。在实验中必须事先从原文关键词中抽取出原文中没有出现的词语，将这些词语作为隐含主题词自动抽取性能的评价依据。

表 2 给出了两种隐含主题词自动抽取方法的结果。通过表 2 可以看出，基于 KNN 或 Citation-KNN 的隐含主题词自动抽取方法具有一定效果。其中基于 Citation-KNN 的隐含主题词自动抽取的查准率高于基于 KNN 的抽取方法，这表明基于

Citation-KNN 算法在分类决策任务中的可靠性要高于 KNN 算法。

同时，通过表 2 还可以看出，两种隐含主题词自动抽取的查准率和召回率都低于 50%。因此我们的下一步工作为寻找提高基于 KNN 或 Citation-KNN 的隐含主题词自动抽取方法质量的方法。

标引模型	P	R	F_1
KNN	0.2586	0.4804	0.3362
Citation-KNN	0.3577	0.4795	0.4097

表 2　隐含主题词自动抽取结果

4 小 结

本文以 KNN 算法为基础，提出基于 Citation-KNN 的隐含主题词自动抽取方法。基于 Citation-KNN 的隐含主题词自动抽取方法是一种懒惰学习算法（Lazy Learning），它利用文本集中与待标引记录相似的文档的关键词，作为待标引记录隐含主题词自动抽取的依据。根据相似文档本身的特征可以进行基于加权方式的 Citation-KNN 的隐含主题词自动抽取。实验结果表明该方法在进行隐含主题词自动抽取这一任务时的有效性。

基于 Citation-KNN 的隐含主题词自动抽取方法存在的问题是，隐含主题词自动抽取的效果强烈依赖于数据集的规模。只有当数据集规模达足够大的情况下，才能充分挖掘出与待标记录内容相似的记录，这样可以保证隐含主题词自动抽取的可靠性。另外，文本间的相似度计算也是一个关键问题，相似度计算的质量对隐含主题词自动抽取的效果有重要影响。

下一步的工作主要包括：获取大量带有关键词的数据集，提高基于 Citation-KNN 的隐含主题词自动抽取方法的可靠性；提出可加可靠的隐含主题词自动抽取的评价方法；探索计算文本间相似度更加可靠的方法。

参 考 文 献

Anjewierden A, Kabel S. 2001. Automatic Indexing of Documents with Ontologies. In: Proceedings of the 13th Belgian/Dutch Conference on Artificial Intelligence (BNAIC-01), Amsterdam, Neteherlands. 23~30.

Baeza-Yates R, Ribeiro-Neto B. 1999. Modern Information Retrieval. New York: Association for Computing Machine (ACM) Press, 27-30.

Chien LF. 1997. PAT-tree-based Keyword Extraction for Chinese Information Retrieval. In: Proceedings of the ACM SIGIR International Conference on Information Retrieval, Philadelphia, USA: ACM Press, 50~59

Cover TM, Hart PE. 1968. Nearest neighbor pattern classification. IEEE Transactions on Information Theory, IT-13 : 21~27

Edmundson H P, Oswald V A. 1959. Automatic Indexing and Abstracting of the Contents of Documents. Planning Research Corp, Document PRC R-126, ASTIA AD No. 231606, Los Angeles. 1~142.

Edmundson H P. 1969. New Methods in Automatic Abstracting Extracting. Journal of the Association for Computing Machinery.16(2): 264~285.

Ercan G, Cicekli I. 2007. Using Lexical Chains for Keyword Extraction. Information Processing and Management, 43(6): 1705~1714.

Frank E, Paynter GW, Witten IH, et al.. 1999. Domain-specific keyphrase extraction. In: Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI-99), California: Morgan Kaufmann, 668~673

Hulth A. 2003. Improved Automatic Keyword Extraction Given More Linguistic Knowledge. In: Proceedings of the 2003 Conference on Emprical Methods in Natural Language Processing, Sapporo, Japan, 216~223.

Lois L E. 1970. Experiments in Automatic Indexing and Extracting. Information Storage and Retrieval, 6: 313~334.

Luhn H P. 1957. A Statistical Approach to Mechanized Encoding and Searching of Literary Information. IBM Journal of Research and Development, 1(4): 309~317

Luhn H P. 1958. The Automatic Creation of Literature Abstracts. IBM Journal of Research and Development. 2(2): 159~165.

Salton G, Wong A, Yang C S. 1975. A Vector Space Model for Automatic Indexing. Communications of ACM, 18(11): 613~620.

Tan P, Steinbach M, Kumar V. 2006. Introduction to Data Mining. Boston: Addison-Wesley, 225.

Tomokiyo T, Hurst M. 2003. A language Model Approach to Keyphrase Extraction. In: Proceedings of the ACL Workshop on Multiword Expressions: Analysis, Acquisition & Treatment, Sapporo, Japan, 33~40.

Turney P D. 1999. Learning to Extract Keyphrases from Text. NRC Technical Report ERB-1057, National Research Council, Canada. 1~43.

Turney PD. 1997. Extraction of Keyphrase from Text: Evaluation of Four Algorithms. Techial Repor ERB-1051, National Research Council, Institute for Information Technology.

Turney PD. 2000. Learning algorithms for keyphrase extraction. Information Retrieval. 2:303~336

Wang J, Zucker J D. 2000. Solving the Multiple-instance Problem: A Lazy Learning Approach. In: Proceedings of 17th International Conference on Machine Learning

(ICML2000). San Francisco: Morgan Kaufmann Publishers, 1119-1125.

Yang Y, Liu X. 1999. A Re-examination of Text Categorization Methods. In: Proceedings of 22nd Annual International ACMSIGIR Conference on Research and Development in Information Retrieval (SIGIR'99), Berkeley, CA, USA, 42~49.

Zhang CZ, Su XN, Zhou DM. 2008. Document Clustering Using Sample Weighting. In: He YX, Xiao GZ, Sun MS eds. Recent Advance of Chinese Computing Technologies Singapore: Chinese and Oriental Languages Information Processing Society, 3: 260-265.

李素建 王厚峰 俞士汶 辛乘胜，2004，关键词自动标引的最大熵模型应用研究，计算机学报，27(9):1192~1197。

Automatic Implicit Semantic Subject Extraction Based on Citation-KNN

Zhang Chengzhi[1, 2] Liu Yao[1] Wang Huilin[1]

1. Institute of Scientific & Technical Information of China, Beijing China, 100038,
2. Department of Information Management, Nanjing University of Science & Technology, Nanjing China, 210094

{zhangchz, liuy, wanghl}istic.ac.cn

Abstract: Currently, the keywords extraction method can only extract words appeared in the article and it cannot extract the implicit semantic subject (ISS). It is a difficult work to extract implicit subject in an article in the task of text mining. As we all know, KNN method is a classic method in machine learning field and is also well used in many other fields. In this paper, we proposed an automatic ISS extraction method based on Citation-KNN method which transformed from the KNN method. Experimental results show that the proposed method can not only improve the precision and recall of keyword extraction, but also extract implicit subject efficiently.

Key Words: Automatic Keyword Extraction; Implicit Semantic Subject; Citation-KNN; Text Mining