关键词自动标引方法综述

厦门大学智能科学系人工仿脑实验室　高　燕

【摘要】本文对关键词提取方法的研究进行了总结。对关键词自动标引方法进行分类梳理，将关键词自动标引方法分为统计分析方法、语言分析方法和人工智能方法三大类；主要介绍了近年比较常用的几种关键词自动标引方法，总结当前关键词自动标引方法存在的问题。
【关键词】关键词自动标引；统计分析方法；语言分析方法；人工智能方法

1.引言

关键词自动标引（Automatic Indexing）技术又可以称为关键词自动抽取（Keywords Extraction）或者术语自动识别（Automatic Term Recognition）。该技术是依靠计算机从文档中选择出反映主题内容的词，可以为用户提供一个简洁的内容摘要，可以说关键词是表达文档内容主题的最小单位，可以使信息定位更加简单便捷。

在当代信息爆炸的社会里，关键词自动标引显得尤为重要。在各个方面都得到广泛的应用，尤其在信息检索、知识挖掘、文本分类、文本聚类等等领域，关键词自动标引更是基础和核心技术。而在相关反馈、自动过滤、事件检测与跟踪等领域，关键词自动标引技术也是起到了比较关键的作用。

目前大多数文档没有标注关键词，而手工标引又费时费力。因此关键词自动标引是一项值得研究的技术。自从1957年，美国人卢恩（H.P.Luhn）提出了基于词频统计的抽词标引法[1]，开始了关键词自动标引技术的探索，到现在的五十多年里，关键词自动提取技术有了很大的发展。本文对现在的关键词自动标引方法进行了系统的分析与梳理。

2.自动提取技术的代表方法总结与分析

自从1957年，卢恩提出了基于词频统计的抽词标引方法之后，几乎平均每五年就会有人提出新的关键词自动抽取方法。根据这些方法所使用的核心理论大致可将它们分为三大类方法：统计分析方法、语言分析方法和人工智能方法。表2.1描述了这三大类自动标引方法的代表方法以及其优劣势。

当然现在的关键词自动提取系统已经没有只靠单一技术来实现的了，基本上都是混合了好几种方法。例如，词性标注与词频统计相结合、词频统计与机器学习方法相结合等等，甚至是好几种方法相结合。而混合方法比单一方法的标引精度要高，但是相对的我们要跟多的考虑几种方法的结合方式。

3.关键词自动抽取方法

3.1 TF·IDF方法

TF·IDF是一种统计方法，用以评估某一个词对于一个文件集或者一个语料库中某一个特定文件的重要程度。TF（term frequency）表示在一份给定的文件里，某一个给定的词语在该文件中出现的次数，为了防止它偏向长的文件，这个数字通常会被归一化。IDF(inverse document frequency)是一个词语普遍重要性的度量，某一特定词的IDF表示一个文件集或者语料库中出现该词的文档

羽校的稳定发展尽管正经历体育的社会化、市场化改革，因为我市经济欠发达，虽然有社会力量开始以赞助形式介入竞技体育后备人才的培养，但是还没有体现出优势，而在未来很长一段时间内，我国竞技体育人才的培养仍主要依托体育系统的三级训练模式，因此，政府对羽校的投入不仅不能退出，而且应该随财政的增长而相应递增。如果一味地要求羽校增强自身的创收能力，势必会影响其训练的宗旨。从羽校近年来的办学情况看，招收自费生已经影响到了后备人才培养的质量。

4.2 积极探索经济转型下多渠道办学形式

由于原有的体制内政策安排已经失去，羽校对少年儿童的吸引力有所下降。现阶段，羽校必须充分考虑儿童少年及其家长的需求，与教育系统进行协作，拓展办学形式，才能在经济转型期增加生源增强办学活力。由于经济的发展，学生家庭条件的改善使得“走训制”更为灵活，离羽校远的学生就住羽校，这样更能吸引有天赋的少年儿童进入到业余训练中来，并且对有天赋的少年儿童进行重点培养，也可就近与湖南城市学院等高校联合组建高水平运动队，构建新型体育后备人才培养，解决羽校学生今后的出路问题，这种做法能够取得良好的效果。同时，羽校还面向社会开展举办俱乐部形式的培训班，以增加自身的创收能力，提高教练员的福利待遇。

4.3 政府给予羽校跨区域招生的政策扶持

针对羽校优质生源短缺的矛盾，政府应打破地区限制，到其它地区甚至省外招生。由于外省市注册运动员难以引进，可以考虑直接到其它地区选材，将具有潜质的苗子引进到羽校参加训练。为此，政府应给予相应的政策扶持，帮助解决运动员户口、入学等问题。由于两所羽校是国家级羽毛球训练基地，对于许多外省市的家长与学生应具有一定的吸引力，这样通过在广大的人才群中选材，招到优质生源，可以在很大程度上缓解生源不足的难题，切实提高后备人才的培养效益。

5.结论与建议

5.1 羽校具有良好的训练条件、可利用的教育资源以及较为稳定的教练员队伍，为培养高水平竞技羽毛球后备人才奠定了良好的物质基础。但是，由于传统体制下体校的优惠政策已经失去，体校吸引力在现阶段有所下降，面临着生源短缺、人才输送链条断裂的现实难题。

5.2 政府应加大对羽校的投入力度，使区域经济优势成为支撑羽校后备人才培养的基本保障。政府应出台相关的政策，通过在教育系统广泛开展业余训练和竞赛，为羽校提供高质量的生源。同时，羽校可在政府政策扶持下进行跨区域招生，挖掘有潜力的苗子。

5.3 羽校应采取多种渠道培养羽毛球后备人才，探索与普通中小学联办课余训练、面向市场举办培训班、或与湖南城市学院联合培养高水平运动队员，既能发现具有羽毛球天赋的学生，培养后备人才；又能推动羽毛球的发展；还能为羽校学生找到今后的出路。

参考文献

[1]黄红泰,伍佰强,胡易,林少娜.广东省羽毛球后备人才的现状与发展[J].内蒙古体育科技(季刊),2009,3:57-58.
[2]谢明正.湖南省竞技羽毛球运动的可持续发展研究[D].湖南师范大学硕士论文,2009年5月.
[3]刘志民.我国竞技体育可持续发展的人力资源研究[J].上海体育学院学报,2000,24(2):6-11.
[4]杨再淮,俞继英等.论竞技体育后备人才资源与可持续发展研究[J].上海体育学院学报,2003,1:1-4.
[5]唐吉平,杨斌等.湖南省羽毛球后备人才管理体制研究[J].中国体育科技,2005,3:11-14.
[6]俞继英.我国竞技体育后备人才培养现状和出路[A].国家体育总局政策法规司.战略抉择——2001年全国体育发展战略研讨会文集[C].内部交流,2001:202-220.

资助项目：湖南省教育厅科研课题（课题号：08C185）。

作者简介：杨清（1971—），男，讲师，研究方向：学校体育学。

数的倒数。TF·IDF主要体现了以下思想：一个词在特定的文档中出现的频率（TF）越高，说明它在区分该文档内容属性方面的能力越强；一个词在文档集中出现的范围越广，说明它区分文档内容属性的能力越弱。TF·IDF算法的经典计算公式为：

$$w_{ij} = tf_{ij} \times idf_j = tf_{ij} \times \log(N / n_j) \quad (1)$$

W_{ij}表示候选词t_j的权值，用来衡量一个词的重要程度；tf_{ij}表示候选词t_j在文档d_i中出现的次数；idf_j表示出现候选词t_j的文档数的倒数；N表示文档集或者语料库中文档总数；n_j表示出现候选词t_j的文档数。

TF·IDF算法是统计关键词自动提取的基本方法，很多其他的方法都是从该方法改进或者变形而来。例如，位置加权法在计算词频的时候，最简单的计算公式可以表示为：

$$w_{ij} = tf_{ij} \times idf_j + tw_{ij} \quad (2)$$

tw_{ij}表示候选词t_j在文档中出现的位置权值。当然还有相对加权、提名加权等都是对于TF或者IDF值计算方法的改进。

3.2 基于词汇链的关键词自动标引方法

基于词汇链的关键词自动标引方法，顾名思义关键技术就是词汇链的构建。这是一种语言分析方法。词汇链是一种词语间语义关系引起的连贯性的外在表现，提供文本结构和主题的重要线索。而词汇链的构建方法有很多：Morris和Hirst[12]提出的用贪婪算法构建词汇链；Barzilay和Elhadad[13]提出用非贪婪算法模型构建词汇链；Silber和McCoy[14]、Gal-leyz和McKeown[15]等也提出了有效的构建词汇链的算法。对于中文文本的词汇链构建，使用最多也是最成熟的方法是利用知网（HowNet）计算语义相似度或者相关度来构建词汇链。

下面所介绍的是一种典型的利用知网计算语义相似度来构建词汇链，进而进行关键词抽取的算法。算法首要任务是构建词汇链，构建词汇链的具体步骤可以如下所述：

1）对文本进行分词和词性标注，取名词作为候选词汇，计算每一候选词的TF，按TF值的大小进行排序；

2）查询知网，获取每一候选词的所有义项，然后确定每一候选词的语义（确定方法有多种，这里不赘述）；

3）将第一个词语语义加入初始词汇链L_0；

4）取下一个候选词，按照顺序依次计算其语义与词汇链L_0中每一词的语义相似度值，将其与规定的阈值比较，如果该值大于规定的阈值θ_0，就将其插入该词汇链，如果比较一轮，所有的语义相似度都小于θ_0，那么创建新的空词汇链L_j，将该候选词语义插入其中；

5）重复步骤4，直到所有的候选词都插入词汇链中。

以该方法构建的词汇链L_j（1≤j≤n，n是候选词个数）实际上是若干个语义相近的词汇的集合。

词汇链构建好以后，我们就可以进行关键词提取。综合考虑词频、候选词出现的位置、词汇链的重要程度等方面，给出一个候选词的权值计算公式。然后计算出候选词的权值，再根据权值对候选词进行排序，取前k（输入值）个词作为关键词。由此，关键词自动标引的具体步骤可以如下所述：

1）按上面1-5的步骤构建词汇链；

2）将TF值、Loc值（位置信息）和词汇链L信息等方面信息整合，得到一个权值计算公式（这个公式可以有很多种变形形式），计算$weight_j$（候选词i的权值）；

3）根据$weight_j$将候选词进行降序排列，取前K个词作为关键词。

这是目前基于语言学分析的关键词自动标引法的主流方法，大概的步骤都如上所述。而权值语义相似度计算方法和候选词权值计算公式有着各种各样的变形形式，而效果也是没有一个标准的评价体系，这里不再赘述。

3.3 KEA方法

KEA系统[16]是由Frank等人提出并实现的关键词提取系统。该系统运用朴素贝叶斯分类器从已经标注了关键词的文档中学习出模型，然后用训练好的模型给新文档抽取关键词。这是一种典型的人工智能方法。

KEA系统主要用到两个特征TF·IDF和词的位置特征。KEA系统通过去除标点、短语识别、去停用词等预处理得到候选词，然后将所有文档的候选词作为候选关键词集合。运用TF·IDF特征和词的位置特征对每一个候选词计算特征，并对得到的特征进行离散数值化处理，得到特征向量。如果候选词在训练集中被标记为关键词，则该候选词就被标记为候选关键词集合中的正例，反之，如果被标记为非关键词，则此候选词就标记为候选关键词集合中的反例。利用分类模型的思想，选取所有的候选词样本作为关键词模型的训练样本。用该训练样本训练贝叶斯分类器，得到关键词提取模型。将此模型用于新文档的关键词抽取。

当对新文档进行关键词抽取时，KEA系统首先识别新文档的候选词，然后计算候选词的特征，根据这些特征计算每一个候选词的权值并排序，最后输出前K（人为设定的需要的关键词个数）个词作为关键词。KEA系统的流程图可以如图3.1所示：

3.4 GenEx方法

GenEx系统[17]是Turney等人在决策树C4.5算法的基础上实现的关键词提取系统。GenEx系统由Genitor和Extractor两个部分组成。Extrator有12个参数，这12个参数需要通过Genitor进行调整，从而使Extractor达到最优。Genitor算法

表2.1 关键词自动标引方法分类分析

方法类型		代表方法	优点	缺点
统计分析方法	一般统计分析法	基于词频[1]、字共现[2]、独立性[3]等方法	简单易行，通用性强	缺乏词义判别，准确性不高
	加权统计分析法	相对加权法[4]、提名加权法、位置加权法等[3]	容易实施	权值不宜确定，准确性不高
	概率统计分析法	基于相关概率的标引方法、2-泊松分布概率标引法等	简单易行	理论模型多，缺少实验验证
	分类判别统计法	BIOSIS分类统计法[5]	标引精度较高	需要有完善的词表，实施困难
语言分析方法	词法分析	词性标注等[6]	简单易行	未登录词的识别与词表的维护问题
	句法分析	短语识别、Churk识别[7]等	可以得到较为专指的短语	准确率不高影响关键词语的识别
	语义分析	词汇链等[8]	能提高质量	目前语义分析准确率不高
	篇章分析	篇章结构分析等[9]	能提高标引质量	对文献结构要求较高，适用范围较窄
人工智能方法	一般机器学习法	NB、最大熵模型、SVM等，相关系统有GenEx、KEA等	不受语种与句型的限制，可以提取出未登录词	对训练集要求较高，存在数据稀疏、过拟合和关键词漏标等问题
	集成学习法	Bagging算法[7]等	标引精度高	需要多个分类器，训练时间较长对训练集要求较高
	专家系统	MedIndEx[10]标引辅助专家系统和JAKS[11]	标引精度高	适用范围较窄，自动学习能力较差，实施困难

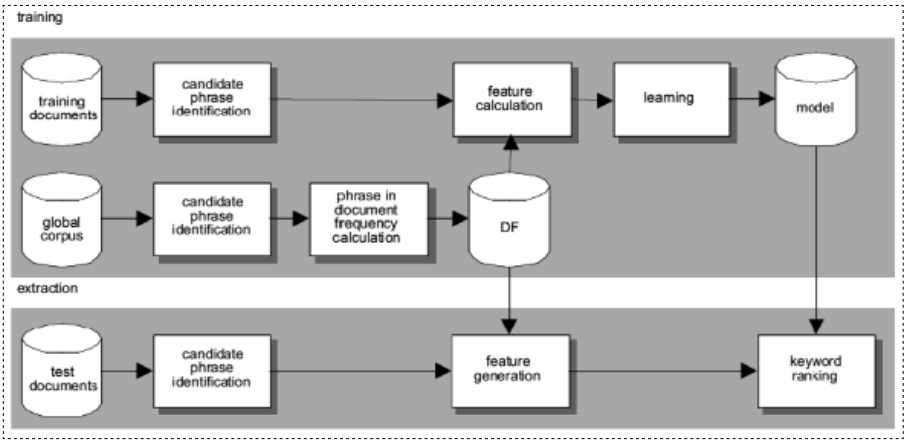

图3.1 KEA的流程图

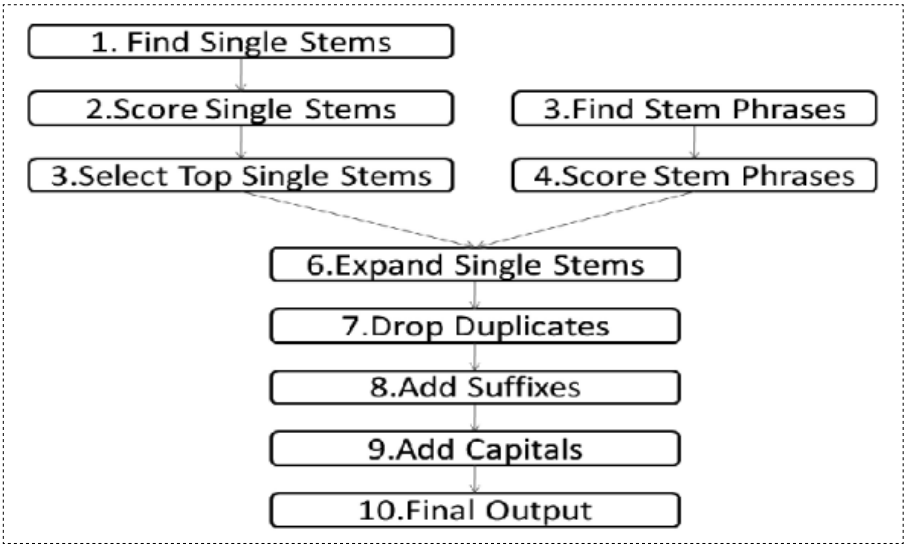

图3.2 Extractor 的流程图

并不死每时每刻都要对Extractor进行调整，而是只要我们确定Extractor的参数达到了最优，我们就可以抛弃Genitor。GenEx系统的Extractor的主要运行流程如图3.2所示。Genitor是一种稳态遗传算法[18]，每一次更新一个个体，产生一个连续变化的群体。在给关键词抽取系统调参的过程中，与Genesis[19, 20]相比，具有更好的稳定性。

4.总结展望

总的来说，关键词自动标引技术从诞生到现在的五十多年来，已经取得了很大的发展。自动标引的准确率有了很大的提高，自动标引的方法也是多种多样，但是关键词的自动标引还有很多的问题。

1）自动标引系统的通用性问题。我们之前介绍的那些关键词自动标引技术总是针对某一个领域或者某一种语言。而与语言无关的关键词自动提取的方法还很少，而且准确率也很低，甚至不超过20%。统计方法的通用性比较强，但是准确率却不高；语言学方法和人工智能的方法准确率较高，但是通用性比较差。怎么解决关键词自动标引系统的通用性问题，将是关键词自动标引技术的一个研究方向。

2）语义分析问题。语言学分析方法中，语义分析方法在当前发展很迅速。但是也存在问题，中文的关键词自动提取中用到的语义分析方法大多数依赖知网，而语义分析仅仅依靠语义词典还是远远不够的。我们需要更好的语义分析知识体系。

3）数据标注瓶颈问题。机器学习方法需要大量的已标注的样本。但是提供尽可能多的已标注样本需要艰苦而缓慢的手工劳动，制约了整个系统的构建。但是，未标注的样本数量很多，而且更接近整个样本空间上的数据分布。如何用少量的已标注样本和大量未标注样本训练出一个好的分类器，将是基于机器学习的关键词自动标引方法的发展方向。

4）知识库的规模问题。专家系统方法将是关键词自动标引的一个发展方向。但是目前知识库的更新慢，跟不上学科的发展。经验证明，开发一个适用的专家系统至少需5人/年。而目前关键词自动标引专家系统与这个要求尚有距离。

总之，虽然关键词自动标引技术多种多样，但由于技术的限制，小规模的实验效果较好，大规模应用的效果还有待提高。关键词自动标引技术距离完全实际应用还有很多问题需要解决，还有很长的路要走。

参考文献

[1]Luhn H P.A Statistical Approach to Mechanized Encoding and Searching of Literary Information[J].IBM Journal of Research and Development,1957,1(4):309-317.
[2]马颖华,王永成,苏贵洋,等.一种基于字同现频率的汉语文本主题抽取方法[J].计算机研究与发展,2004,40(6):874-878.
[3]Ednundson H P.New Methods in Automatic Abstracting Extracting[J].Journal of the Association for Computing Machinery,1969,16(2):264-285.
[4]Ednundson H P,Oswald V A.Automatic Indexing and Abstracting of Contents of Documents[R].Planning Research Corp,Document PRCR-126,ASTRIAAD No.231606,Los Angeles,1959:1-142.
[5]Vledutz-Stokolov,N.Concept Recognition in an Automatic Text Processing System for the Life Science[J].Journal of the American Society for Information Scence,1987(4):269-297.
[6]韩客松,王永成.中文全文标引的主题词标引和主题概念标引方法[J].情报学报,2001,20(2):212-216.
[7]Hulth A.Improved Automatic Keyword Extraction Given More Linguistic Knowedge[A].In Proceedings of the 2003 Conference on Emprical Methods in Natural Language Processing[C].Sapporo,Japan,2003:216-223.
[8]索红光,刘玉树,曹淑英.一种基于词汇链的关键词抽取方法[J].中文信息学报,2006,20(6):25-30.
[9]Salton G,Buckley C.Automatic Text Structuring and Retrieval:Experiments in Automatic Encyclopedia Searching[A].In:Proceedings of the Fourteenth SIGIR Conference[C].New York:ACM,1991:21-30.
[10]Humphrey,S1M.MedlndEx System:Medical Indexing Expert System[J].Information Processing and Management,1986(1):73-88.
[11]Driscoll,J1R1,et al.The Operation and Performance of an Artificially Intelligent Keywording System[J].Information Processing and Management,1991(1):43-54.
[12]Morris J,Hirst G.Lexical Cohesion Computed by Thesaural relations as an Indicator of the Structure of Text.Computational Linguistics,1991,17(1):2-48.
[13]R.Barzilay,M.Elhadad.Using Lexical Chains for Rext Summarization.Proceedings of the Intelligent Scalable Text Summarization Workshop(ISTS-97),ACL,Madrid,Spain,pages:10-18.
[14]Silber H,McCoy G.Efficiently Computed Lexical Chains as an Intermediate Representation for Automatic Text Summarization.Computational Linguistics,2002(4):487-496.
[15]Galley M,McKeown K.Improving Word Sense Disambiguation in Lexical Chaining.Proc of the 18th International Joint Conference on Artificial Intelligence.Acapulco,Mexico,2003:1486-1488.
[16]Turney P D.Learning to Extract Keyphrases from Text[R].NRC Technical Report ERB-1057,National Research Council Canada.1999:1-43.
[17]Witten I H,Paynter G W,Frank E,et al.KEA:Practical Automatic Keyphrase Extraction.Proc of the 4thACM Conference on Digital Libraries.Berkeley,USA,1999:668-673.
[18]Whitley D.The GENITOR algorithm and selective pressure.Proceedings of the Third International Conference on Genetic Algorithms(ICGA-89),1989:116-121.California:Morgan Kaufmann.
[19]Grefenstette J J.A user's guide to GENESIS.Technical Report CS-83-11,Computer Science Department,Vanderbilt University,1983.
[20]Grefenstette J J.Optimization of control parameters for genetic algorithms.IEEE Transactions on Systems,Man,and Cybernetics,1986,16,122-128.