情报工程 ISSN: 2095-915X
第1卷 第5期 25-31 2015 年 10 月
TECHNOLOGY INTELLIGENCE ENGINEERING
ISSN: 2095-915X Vol.1 No.5 25-31 Oct 2015

doi:10.3772/j.issn.2095-915x.2015.05.004

基于 TValue 融合领域度的术语抽取法 [1]

杨雅娜 [1]，刘胜奇 [2]

（1. 中国邮政储蓄银行 北京 100070；2. 中国专利信息中心 北京 100088）

摘要：提出 ATValue(Advanced TValue and Fieldhood Integration) 术语抽取法。为提高术语抽取质量，在 TValue 五属性的基础上，提出领域度。通过相关性分析获得六属性组合值 AValue，最后识别 AValue 高于术语可信度的词串来选择候选术语。能源行业的实验结果表明：ATValue 术语抽取法的 F 值约比 TValue 术语抽取法高出 2 个百分点，原因在于 ATValue 的领域度测算了词串中各种单词对领域的贡献。

关键词：术语抽取，术语识别，数据挖掘，领域度

中图分类号：TP391.1, G306.0

Automatic Term Extraction Based on Advanced TValue and Fieldhood Integration

YANG Yana [1], LIU Shengqi [2]

(1. Postal Savings Bank of China, Beijing, 100070, China; 2. China Patent Information Center, Beijing, 100088, China)

Abstract: It proposes an automatic term extraction based on ATValue (advanced TValue and fieldhood integration). In order to increase the quality of term extraction, it puts forward the degree of fieldhood based on the five attributes of TValue. The value of AValue is computed by the six attributes of the strings based on

作者简介：杨雅娜和刘胜奇与第一作者贡献相同，为并列第一作者。杨雅娜（1986-），本科，研究方向：专利分析、银行风险管理。刘胜奇（1978-），博士，高级工程师，研究方向：知识管理、可视化、创新管理，shengqiliu@126.com。

multiplication of probability after their correlations are analyzed. It gains the candidate terms by the analysis of the strings whose value of AValue is more than the pre-defined confidence threshold. The simulation results of term extraction in energy industry show that the F-score of automatic term extraction based on ATValue is about 2% higher than that based on TValue, because it measures the score of importance of compound words by the degree of fieldhood of ATValue.

Keywords: Term Extraction, Term Recognition, Data Mining, Fieldhood

1 引言

术语抽取对于信息检索 [1]、大数据管理 [2]、机器翻译 [3]、客户关系管理 [4]、自动文摘 [5]、技术空白分析 [6]、替代技术分析 [7] 等，具有很大的应用价值。

历史研究中，经典的 C/NC-value[8]、名词计分法 [9] 及其衍生方法，召回率不超过 85%[9]，准确率不超过 76%[8]。召回率不高，主要因为所用词法规则为词性固定组合，覆盖度低，若仅考虑首尾词性规则，可使召回率大大提高。准确率不高的核心原因有二：一是不完整词串的干扰，二是停用词串 (非术语) 的误判。C/NC-value 中的子串修正 [8]，实质涉及词串独立存在的概率，显著减少干扰，但忽略了中间最短母串 (左右两边各加一词构成的词串) 的影响。TValue[10] 用中间最短母串修正母串对子串的影响，提出词串独立度。名词计分法 [9] 实质计算词串在特定领域的停用情况，但仅考虑名词术语，忽略单名词非临近词影响和复合名词外其他词影响，方法有效却极端。TValue[10] 综合考虑各种词性词串的停用情况，提出词串停用度。此外，TValue[10] 有效组合了首尾词性度、词长度、停用度、独立度和重要度，实验表明 TValue 术语抽取质量优于同类方法。

然而，TValue[10] 术语抽取法的实验准确率约为 84.08%，召回率约为 94.49%，存在改进提高的空间。下面，进一步分析名词计分法 [9]：

(1) TValue[10] 已讨论假设“每个名词对领域的贡献相同”不成立，例如“电”领域中的“电”，对领域的贡献度远超其他词。由此，词串内每个单词对领域存在不同程度的贡献。为测算词串内单词对领域的贡献大小，本文提出内领域度。

(2) TValue[10] 已说明假设“忽略单名词非临近词影响和复合名词外其他词影响”不成立。对单名词来说，临近名词 (前后名词) 对单名词的领域影响很大，其他词语对单名词的领域影响也存在；同理，复合名词内含名词对其有领域贡献，外部的其他词语对它也有领域贡献。为测算词串外其他词串对该词串的领域影响，本文提出外领域度。

(3)名词计分法 [9] 中的实验表明质量最佳时，名词计分法 (a=1 时) 的公式可简化为几何均值或算术均值。本文提出的内、外领域度，本质是一种概率可能性，不能通过均值计算，可通过概率加法法则进一步组合。由此，本文提出领域度。

综上，基于 TValue 术语抽取法中词串的五个属性 (首尾词性度、词长度、停用度、独立度和重要度)，结合本文提出的领域度，本文提出 ATValue(Advanced TValue and Fieldhood Integration) 术语抽取法，计算词串的六属性组合值 AValue，若词串的 AValue 高于 λ (λ 为术语可信阈值)，则为候选术语。

2 ATValue 术语抽取法

2.1 内领域度

定义 1：内领域度测量词串本身所含的领域知识。通过概率加法法则，用领域种子的内领域度，计算内领域度如下：

$$d_1(STR)=P\left(\sum_{m=1}^{M}SEED_m\right)$$

P 表示基于概率加法法则计算，$SEED_m$ 表示词串的第 m 个领域种子的内领域度，M 表示词串共有 M 个领域种子。领域种子的内领域度计算如下：

$$SEED_m=f(SEED_m)/\Omega$$

$f(SEED_m)$ 表示术语库中含有领域种子 $SEED_m$ 的术语个数；Ω 表示术语库中术语总数。

例如：对于词串 STR= 消音 /v 装置 /n，只有一个领域种子“装置”，则内领域度为：

d_1(消音 /v 装置 /n)=1449/76869 ≈ 0.01885。

现记 $SEED(STR)$ 为词串的领域种子集；m 为 $SEED(STR)$ 的领域种子个数；采用递归算法的词串内领域度算法 d_1 如下：

Function $d1(STR, m)$

IF m==0

// 词串 STR 中没有领域种子

$d1(STR, m)=0$

ELSE IF m==1

// 词串 STR 只含一个领域种子

$d1(STR, m)$= 领域种子内领域度

ELSE

// 若词串含有多个领域种子，据概率加法法则计算，词串内领域度 = 去掉一个领域种子的词串内领域度 + 去掉的领域种子的内领域度 − 去掉的领域种子的内领域度 × 去掉一个领域种子的词串内领域度

$d1(STR, m)= d1(STR, m-1) + SEED(STR)[m]$ 的内领域度 $-d1(STR, m-1)\times SEED(STR)[m]$ 的内领域度

END IF

END Function

2.2 外领域度

定义 2：外领域度测量其他词串对该词串的领域贡献。基于文档 k 中同句的其他词串的内领域度，计算外领域度如下：

$$d_{2k}(STR)=P\left(\sum_{n=1}^{N_k}\frac{d_1(STR_{kn})}{L_{kn}+1}\right)$$

P 表示基于概率加法法则计算，$d_1(STR_{kn})$ 表示文档 k 中词串 STR 共句的第 n 个词串的内领域度； 表示文档 k 中第 n 个共句词串与 STR 之间间隔的词串个数，为防止分母为 0，所以用 $L_{kn}+1$ 作为分母；N_k 表示文档 k 中与 STR 共句的有 N 个词串。

例如：对于词串 STR= 消音 /v 装置 /n，在 CN1664348 的专利名称及摘要中，“STR= 消音 /v 装置 /n”只出现在专利名称及摘要的第一句中，共句的词串情况如下表 (表中相同的共句词串表

表 1 共句词串情况

共句词串	共句词串内领域度	$L_{kn}+1$
液化 /vn 石油 /n 喷射 /v	0.10467	6
包括 /v 一 /m 个 /q 液化 /vn 石油气 /n	0.39093	2
石油 /n 喷射 /v 车辆 /n	0.05539	5
车辆 /n 中 /f 的 /ude1 燃 /v	0.03788	2
液化 /vn 石油 /n 喷射 /v 车辆 /n	0.10467	5
液化 /vn 石油气 /n	0.39093	5
发明公开 /n 一 /m 种 /q 液化 /v	0.05217	11
燃 /v	0.03788	2
发明公开 /n 一 /m 种 /q 液化 /vn 石油 /n 喷射 /v	0.10467	9
液化 /vn	0.05217	8
液化 /vn	0.05217	11
车辆 /n 中 /f 的 /ude1 燃 /vg 油泵 /n 消音 /v	0.09117	1
石油 /n	0.05539	7
石油 /n	0.05539	10
包括 /v 一 /m 个 /q 液化 /v	0.05217	2
石油气 /n	0.35741	6
燃 /vg 油泵 /n	0.09117	2
石油 /n 喷射 /v	0.05539	6
液化 /vn 石油 /n 喷射 /v	0.10467	6

示二者在多个句子中共现）：

则词串“消音/v 装置/n”在文档 CN1664348 中的外领域度为：

d_{2k}（消音/v 装置/n）

$$= P\left(\sum_{n=1}^{19} \frac{d_1((\text{消音}/v\ \text{装置}/\text{n})_{kn})}{L_{kn}+1}\right)$$

≈ 0.496043

现记 *CO_S(STR)* 为有词串 *STR* 的文档 *k* 中的共句词串集；n 为 *CO_S(STR)* 中的词串个数；*L(STR, k, n)* 为词串 *STR* 与第 *n* 个共句词串之间的词串个数，若之间还存在非词串的字符组合，则非词串的字符组合按分词后的单词计算，非词串的一个单词计为一个词串（其内领域度记为 0）。采用递归算法的词串 *STR* 的外领域度算法 *d2* 如下：

Function *d2(STR, k, n)*

IF *n*==0

//CO_S(STR) 中没有共句词串

d2(STR, k, n)=0

ELSE IF n==1

//CO_S(STR) 中只有一个共句词串

d2(STR, k, n)= d1(CO_S(STR)[1]) / *(L(STR, k, n)+1)*

ELSE

// 若词串有多个共句词串，根据概率加法法则计算，词串外领域度 = 去掉一个共句词串的词串外领域度 + 去掉的共句词串的内领域度 /(L_{kn}+1) − 去掉的共句词串的内领域度 /(L_{kn}+1) × 去掉一个共句词串的词串外领域度

d2(STR, k, n)= d2(STR, k, n-1) + d1(CO_S(STR)[n]) / (L(STR, k, n)+1) - d2(STR, k, n-1)×d1(CO_S(STR)[n]) / (L(STR, k, n)+1)

END IF

END Function

2.3 领域度

定义 3：领域度，表示在文档 *k* 中，词串对领域知识的表达能力。由内领域度和外领域度，组合领域度计算如下：

$$d_k(STR) = d_1(STR) + d_{2k}(STR) - d_1(STR)\cdot d_{2k}(STR)$$

$d_1(STR)$ 表示词串 *STR* 的内领域度；$d_{2k}(STR)$ 表示词串 *STR* 在文档 *k* 中的外领域度。

例如：词串“消音/v 装置/n”在文档 CN1664348 中的领域度计算为：

d_k（消音/v 装置/n）

$=d_1$（消音/v 装置/n）$+d_{2k}$（消音/v 装置/n）$-d_1$（消音/v 装置/n）$\times d_{2k}$（消音/v 装置/n）

$\approx 0.01885+0.496043-0.01885\times 0.496043$

≈ 0.505543

2.4 AValue

AValue 衡量一个词串能构成术语的程度，由词串的六个属性（首尾词性度 p、词长度 l、独立度 i、停用度 s、重要度 tk 和领域度 dk）组合而成。

为获取六个属性的组合规则，在训练术语库中，计算术语六个属性的相关系数如表 2：

表 2 六个属性的相关系数

	s	*l*	*i*	*p*	*tk*	*dk*	*log(2, tk)*	s^2
s	1							
l	0.35967	1						
i	0.02648	0.09433	1					
p	−0.13897	−0.24929	−0.07278	1				
tk	0.10292	0.10618	0.05036	−0.02219	1			
dk	0.05020	0.09829	0.03112	−0.05176	0.05692	1		
log(2,tk)	0.07715	0.04406	0.00411	−0.00977	0.54416	0.00944	1	
s^2	0.78118	0.08926	0.01837	−0.07957	0.08346	0.03032	0.10477	1

从表 2 的相关系数可以看出，词串六个属性之间相关性较弱，可认为基本独立，所以考虑采用乘法法则进行六个属性组合分析。

另外，训练术语库中术语的六个属性值还表明：

（1）中文词串独立度存在 0 的可能，为防止 0 的极端影响，加 0.0001，计为 $i(s)+0.0001$。

（2）词串停用度的平方与其他属性的关系更弱，其倒数可表示词串在文档中可构成术语的程度，计为 $1/s^2(s)$。

（3）词串重要度经过对数变换后，与其他属性的关系更弱，所以组合时可考虑用 $log(2,tk(s))$ 代替 $tk(s)$。

由此，词串 s 在文档 k 中的 AValue 计算为：

$$AValue_k(s)=\left(\frac{l(s)\cdot(i(s)+0.0001)\cdot p(s)\cdot d_k(s)}{s^2(s)}\right)\cdot\log(2,t_k(s))$$

同理在训练术语库中，统计英文术语 6 个属性的相关系数，虽与上表略有差异，但分析发现词串 s 在英文文档 k 中的 AValue 计算式还是上式为佳。

3 实证分析

下面以能源行业专利为例，完成 ATValue 术语抽取法的实证分析。

3.1 语料准备

（1）专利样本库

从国家知识产权局专利库[2]中，获取能源行业的部分中英文专利 9358 个，经过人工筛选后约 8710 个，包括标题、摘要。随机抽取 100 个有中英文摘要的中英文专利，组成专利样本库。

（2）训练术语库

在互联网中，共获得中英文对照的能源术语术语 76869 对，分词[3 4]，后建成训练术语库。

3.2 结果评价方法

为便于估算 ATValue 术语抽取法的有效性，采用以下评价方法：

准确率[11]= 样本中抽取的正确术语数 CT/ 样本中抽取的术语数 ET

召回率[11]= 样本中抽取的正确术语数 CT/ 样本中人工标注的术语总数 LT

F 值[11]=2*准确率*召回率/(准确率*召回率)

3.3 中文术语抽取结果评价

人工处理专利样本库中的 100 个中文专利样本，获得人工标注的中文术语总数 LT=4121 条。通过术语可信阈值估计[10]，设定 λ =0.0000001，ATValue 术语抽取法抽取的术语数 ET=4242 条。再分别用 C-value、NC-value、Gensen Web(基于名词计分法)、TValue(λ 设为 0.0000005) 对中文专利样本进行术语抽取，然后分别取排名前 4442 的候选术语。若抽取的术语与人工标注的术语一致，则为样本中抽取的正确术语。中文术语抽取对比评价结果如表 3：

表 3 中文术语抽取结果评价

方法	ET	CT	LT	准确率	召回率	F 值
C-value	TOP4442	2681	4121	60.36%	65.06%	62.62%
NC-value	TOP4442	2873	4121	64.68%	69.72%	67.11%
Gensen Web	1426	1085	4121	76.09%	26.33%	39.12%
TValue	4440	3796	4121	85.50%	92.11%	88.68%
ATValue	4442	3904	4121	87.89%	94.73%	91.18%

2 http://www.pss-system.gov.cn/sipopublicsearch/ search/searchHomeIndex. shtml
3 http://ictclas.org
4 http://nlp.stanford.edu/software/tagger.shtml

从表 3 可以看出：

（1）C-value、NC-value、Gensen Web、TValue 的准确率均略高于文献 [10] 中的相应数据，说明术语抽取的准确率和样本的选择略有关系。

（2）C-value、NC-value、TValue 的召回率略低于文献 [10] 中的数据，也是由于样本不同。

（3）ATValue 的中文术语抽取准确率约为 87.89%，召回率约为 94.73%，F 值约为 91.18%，均比 TValue 高出 2 个百分点。

3.4 英文术语抽取结果评价

人工处理专利样本库中的 100 个英文专利样本，获得人工标注的中文术语总数 LT=4182 条。通过术语可信阈值估计 [10]，设定 λ =0.00000015，ATValue 术语抽取法抽取的术语数 ET=4383 条。再分别用 C-value、NC-value、Gensen Web（基于名词计分法）、TValue（λ 设为 0.0000007）对英文专利样本进行术语抽取，然后分别取排名前 4383 的候选术语。若抽取的术语与人工标注的术语一致，则为样本中抽取的正确术语。英文术语抽取对比评价结果如表 4：

表 4 英文术语抽取结果评价

方法	ET	CT	LT	准确率	召回率	F 值
C-value	TOP4383	2685	4182	61.26%	64.20%	62.70%
NC-value	TOP4383	3164	4182	72.19%	75.66%	73.88%
Gensen Web	1485	1086	4182	73.13%	25.97%	38.33%
TValue	4333	3911	4182	90.26%	93.52%	91.86%
ATValue	4383	4045	4182	92.29%	96.72%	94.45%

从表 4 可以看出：

（1）ATValue、C-value、NC-value、TValue 的英文术语抽取准确率均高于中文术语，原因有二：一是英文专利中单词本身以空格间隔，英文词法分析器的分词、词性标注效果，都好于中文词法分析器；二是英文专利摘要的撰写质量高于中文专利摘要，如“激光加工”专利的英文摘要，大量使用一个单词表达一个术语，所以英文术语抽取准确率和召回率都不错，而中文摘要存在大量不规则的表达，导致词法分析产生大量错误。

（2）Gensen Web 的英文术语抽取准确率反而低于中文术语，可能因为 Gensen Web 基于名词计分法，而名词计分法对单词间无间隔的中文、日文有特效，但不擅长处理英文。

（3）ATValue 的英文术语抽取准确率约为 92.29%，召回率约为 96.72%，F 值约为 94.45%，也比 TValue 高出 2 个多百分点。原因在于：TValue 不考虑词串中的单词对领域的贡献；而 ATValue 的领域度，考虑了词串中各种单词对领域的贡献，使得抽取的术语更接近真实的领域术语。

结论

本文提出的 ATValue 术语抽取法，在 TValue 五属性的基础上，新增属性领域度，并通过相关度计算有效组合六属性。

能源行业专利样本库实验结果，表明 ATValue 术语抽取法的质量高于 C-value、NC-value、Gensen Web（基于名词计分法）、TValue。

5 http://gensen.dl.itc.u-tokyo.ac.jp/gensenweb_eng.html

本文实验验证只采用能源行业专利数据，今后将进一步扩展方法、开展多领域实验。进一步研究可考虑一下方向：

（1）新增词串属性，自动获取词串属性参数、由属性相关性自动构造属性组合函数。

（2）将实验分析扩展到其他语料，如论文、新闻、评论等。

参考文献

[1]YANAGIHORI K, TANAKA K, TSUDA K. Improvement of Terminology Extraction Method for Specific Patent Search [J]. Procedia Computer Science, 2014, 35: 879-885.

[2]PÉREZ M J M, RIZZO C R. Automatic Access to Legal Terminology Applying two Different Automatic Term Recognition Methods. CILC2013 [J]. Procedia - Social and Behavioral Sciences, 2013, 95:455-463.

[3] 李丽双 , 党延忠 , 张婧 , 等 . 基于条件随机场的汽车领域术语抽取 [J]. 大连理工大学学报 , 2013, 53(2): 267-272.

[4]ITTOO A, BOUMA G. Term Extraction from Sparse, Ungrammatical Domain-Specific Documents [J]. Expert Systems with Applications, 2013, 40(7): 2530-2540.

[5] 熊李艳 , 谭龙 , 钟茂生 . 基于有效词频的改进 C-Value 自动术语抽取方法 [J]. 现代图书情报技术 , 2013, 29(9): 54-59.

[6]SON C, SUH Y, JEON J, et al. Development of a GTM-based Patent Map for Identifying Patent Vacuums [J]. Expert Systems with Applications, 2012, 39(3): 2489-2500.

[7] 娄岩 , 张赏 , 黄鲁成 , 等 . 基于专利分析的替代性技术识别研究 [J]. 情报杂志 , 2014, 33(9): 27-32.

[8]KATERINA T F, SOPHIA A, HIDEKI M. Automatic Recognition of Multi-word Terms: the C-value/NC-value Method [J]. Intemational Joumal on Digital Libraries, 2000, 3(2): 115- 130.

[9]HIROSHI N, TATSUNORI M. Automatic Term Recognition Based on Statistics of Compound words and their Components [J]. Journal of Terminology, 2003, 9(2): 201-219.

[10] 刘胜奇 , 朱东华 . TValue 术语抽取法 [J]. 情报学报 , 2013, 32(11): 1164-1173.

[11]ITTOO A, BOUMA G. Term Extraction from Sparse, Ungrammatical Domain-Specific Documents [J]. Expert Systems with Applications, 2013, 40(7): 2530-2540.