知识组织与知识管理

基于改进 C – value 方法的中文术语抽取

胡阿沛　张　静　刘俊丽
(中国科学技术信息研究所　北京 100038)

【摘要】提出一种改进 C – value 的术语抽取方法,即 IC – value 方法。利用停用词对文本进行预处理后,采用一种基于串频统计的抽取算法提取候选术语;对候选术语进行语言规则过滤;从逆文档频率、破碎子串和术语长度三个方面改进 C – value 方法得到 IC – value 方法,并用来计算候选术语的术语度。以 1 000 篇乙型肝炎相关论文摘要进行实证研究,结果证明 IC – value 方法在准确率和召回率方面都要优于 C – value、TF – IDF 和 V – value,有较强的长术语发现能力,且识别破碎子串的效果十分明显。

【关键词】术语抽取　串频统计　语言规则　术语度
【分类号】TP391.1

Chinese Term Extraction Based on Improved C – value Method

Hu Apei　Zhang Jing　Liu Junli
(Institute of Scientific & Technical Information of China, Beijing 100038, China)

【Abstract】An improved C – value term extraction method is introduced in the paper. Firstly, the domain – specific text corpora is preprocessed by stop word list. Secondly, a term extraction algorithm based on the co – occurrence frequency of multi – character is applied to get candidate terms. Lastly, term selection is completed based on termhood computed by IC – value which is the improvement of C – value in terms of inverse document frequency, meaningless substring and term length. Empirical study is conducted based on 1 000 abstracts of articles about Hepatitis B. The results indicate the proposed IC – value is much better than C – value, TF – IDF and V – value in both precision and recall. And IC – value also has good performance in long term extraction and it is very effective in filtering meaningless substring.

【Keywords】Term extraction　Statistics of string frequency　Linguistical rules　Termhood

1　引　言

术语是某种语言中专门指称某一专业知识活动领域一般(具体或者抽象)理论概念的词汇单位,它是专业领域知识系统中的重要组成部分,集中体现和承载了一个学科领域的核心知识[1,2]。当今科学技术高速发展,各个学科领域的新知识不断涌现,与之相应的是学科领域内术语的层出不穷。通过术语可以快速了解各个学科的发展动态,对于科学研究甚至国家发展规划具有重要意义。但是依靠人工收集不断涌现的新术语费时费力,跟不上更新速度,然而,利用计算机实现术语自动抽取可以解决这一问题。另外,术语自动抽取可以辅助编纂专业词典[3],同时也是信息检索、文本挖掘、机器翻译[4]等的重要部分。

归纳现有的术语自动抽取方法,主要分为以下 4 类。

收稿日期:2013 – 01 – 04
收修改稿日期:2013 – 02 – 15

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

(1) 基于词典的方法。该方法通过与领域词典中的词匹配抽取术语,简单易行,但存在大量术语变体不易识别、术语不断更新使得领域词典不易维护等问题[5]。

(2) 基于语言学规则的方法。该方法基于以下假设,即术语作为一个独立的语言单位,其语言结构也应该是稳固的[6]。也就是说术语需要满足一定的语法结构。语言学规则的方法通过构建语言规则来识别术语。但语言学规则难以发现,费时费力。同时,语言学规则识别术语时会产生大量噪音。并且不同领域的术语语言学规则也有所差异,导致可移植性差。

(3) 基于统计学的方法。该方法以统计学理论为基础,利用术语在语料中的统计属性来识别其中潜在的术语。常用的一些统计学方法有串频统计[7]、互信息[8,9]、Log - likelihood[10]等。统计学的方法受专业领域的限制小,移植性良好。但利用统计学方法提取术语会存在无意义的字串组合(如公共破碎字串[11])、常用词语(非术语)等噪音。

(4) 基于机器学习的方法。如岑咏华等[12]利用隐马尔科夫模型进行中文术语识别研究。但这些模型是监督型的方法,需要一个适当的训练集对模型参数进行训练,而获得合适的训练集要耗费大量时间和人力,并且不同的领域训练集也不同,导致训练出来的模型可移植性差。

实际的术语抽取往往将以上方法有选择性地结合在一起。比如为了充分利用语言学方法和统计学方法两者的优点,常常使用规则与统计相结合的方法。

本文使用统计与语言规则结合的方法获取候选术语集,并提出一种改进的 C - value 方法,即 IC - value 方法,来计算候选术语的术语度值。利用停用词表对文本进行预处理,在此基础上,采用一种串频统计的方法提取文本中的重复字符串,把这些字符串作为候选术语;采取逆向的方法,利用候选术语的词性将一些明显不能作为术语的候选术语过滤掉;从逆文档频率、公共破碎子串和术语长度三个方面改进了 Frantzi 等[13]提出的 C - value 方法,用于计算候选术语的术语度(Termhood),并取一定的阈值过滤候选术语。该方法既不需要专业领域的词典,也不需要训练模型和总结语法规则。术语抽取流程如图 1 所示。

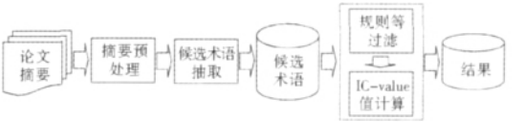

图 1　术语抽取流程

2　候选术语的获取

首先利用停用词表预处理文本,这样可以提高术语识别的效率和正确率。之后,采用一种串频统计的方法提取文本中的重复字符串作为候选术语。

2.1　停用词选取

本文选取乙型肝炎相关论文的摘要作为实验数据。从互联网上获得一般停用词[14];通过阅读 50 篇论文摘要,人工确定与该领域相关的停用词,同时剔除一些与该领域无关的一般停用词。由于医学文献的特殊性,有大量缩写和其他非中文字符,因此本文也将字母和阿拉伯数字作为停用词使用。

2.2　基于串频统计的候选术语抽取

本文采用的候选术语抽取算法基于以下假设:某篇专业领域文档中重复出现的字符串很有可能是该领域的专业术语[7]。比如在一篇有关乙型肝炎的摘要中,"乙型肝炎病毒"一词多次出现。本文采用的候选术语抽取算法说明如下:

候选术语抽取算法的过程类似于寻找最大频繁项集的过程。若将文本看作字符串 $str = a_1 a_2 \cdots a_n (n \geq 2)$,其中 $a_i (1 \leq i \leq n)$ 表示字符串中的一个字符,则候选术语抽取即是在字符串 str 中寻找出现频次满足阈值的最长子串 $substr = a_i \cdots a_j (1 \leq i \leq j \leq n)$。最长子串的概念如下:若某满足阈值要求的子串与相邻字符合并得到的子串不能满足阈值要求,则称其为最长子串。结合本问题的特殊约束条件,可将问题转化为寻找最大频繁项集(相应于最长字串的概念,与通常意义上的最大频繁项集有所差异)的过程如下:

设 $t = \{a_1, a_2, \cdots a_n\}$ 表示文本,a_i 项表示字符串中的字符,t 中的任何连续项 $\{a_i, \cdots a_j\} (1 \leq i \leq j \leq n)$ 构成一个项集。将项集 $\{a_i, \cdots a_j\}$ 的支持度设为其在 t 中的出现频次,即 $support(\{a_i \cdots a_j\}) = freq(\{a_i \cdots a_j\})$,若 $support(\{a_i \cdots a_j\}) \geq \sigma$,则称项集 $\{a_i, \cdots a_j\}$ 为频繁项集,其中 σ 是设定的阈值,即重复的频次。候选术语抽取即转换为在 t 中寻找满足 $support(\{a_i \cdots a_j\}) \geq \sigma$ 的

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

最大频繁项集。在算法实现上可参考 Apriori 算法。

值得注意的是该算法会提取出现频次大于其父串的重复子串。例如,若重复子串“乙型肝炎病毒”在文本中的出现频次大于父重复子串“慢性乙型肝炎病毒”的出现频次,则两者都将被提取,否则,仅提取“慢性乙型肝炎病毒”。因此,该算法可以提取嵌套术语。

3 术语选择

3.1 语言学规则过滤

由于术语词性构成规则很难总结,并且几乎每种规则都会产生噪音。因此,本文采取逆向的方法,利用候选术语的词性将一些明显不能作为术语的候选术语过滤掉。先采用中国科学院计算技术研究所的汉语词法分析系统 ICTCLAS 对候选术语进行分词及词性标注;再对经过分词及词性标注的候选术语总结明显不能作为术语的语言学规则,如表 1 所示:

表 1 非术语词性构成规则

编号	规则描述
1	结尾是数词(m)、方位词(f)、介词(p)和连接词(k)
2	单个词,且非名词
3	包含分隔符(w)

3.2 术语度计算(IC – value)

候选术语中一般包括有普通词语搭配、无意义的字串和术语,要从中选择出正确的术语,简单的有频次排序选择法[8],该方法虽简单易行,但是仅考虑候选术语的出现频次而没有考虑候选术语的长度以及文档频率;使用领域相减[2]的方法可以过滤掉一般词语,但需要一个对照语料库,且存在一般词语识别能力不足等问题;基于 TF – IDF 方法及其变形形式的术语选择也是常用的方法[11]。以上方法都没有考虑到嵌套术语和候选术语长度的问题。

相比之下,C – value 方法虽然考虑了术语长度和嵌套术语,但也存在一些问题。目前,已有不少人从不同的角度对 C – value 方法进行改进。许德山等[13]结合上下文环境在 C – value 基础上进行改进,提出 V – value 方法,在一定程度上提高了术语识别的准确率。还有人将领域类别信息[16]、文档频率[17]等融入 C – value 值计算中。以上改进方法没有考虑到逆文档频率、公共破碎子串以及提高长术语发现能力的特殊意义。

本文从逆文档频率、公共破碎子串和术语长度三个方面改进 C – value 方法,具体说明如下:

(1) C – value 方法不能有效过滤一些出现频次很高的普通词汇。按 C – value 方法计算术语度,则出现频次高的普通词汇会获得较高的 C – value 值,而无法得到有效过滤,因此本文在 C – value 值计算中融入逆文档频率,降低高频次普通词汇的术语度值。

(2) C – value 方法不能很好地区别公共破碎子串与嵌套术语。例如,“病毒性肝炎肝衰竭”与“病毒性乙型肝炎”都是词串“病毒性”的父串,“病毒性”为公共破碎子串,而并非嵌套术语。“细胞免疫”的父串有“细胞免疫应答”、“细胞免疫功能”等,“细胞免疫”为术语。事实上,若某一子串是公共破碎子串,就不应该独立出现;而若某一子串是术语,就应该会独立出现。因此,为了有效区分公共破碎子串(如“病毒性”)与术语(如“细胞免疫”),在计算中将 C – value 中的 $f(a) - \sum_{b\in T_a} f(b)/P(T_a)$ 改为 $f(a) - \sum_{b\in T_a} f(b)$,即计算子串独立出现的频次。

(3) 短术语多为如细胞、基因、乙肝、血清和病毒的术语,在领域内具有相对普遍的意义。该类术语可以与其他词组合成更具针对性的长术语,比如淋巴细胞、肝炎病毒等。长术语一般具有更为特定的含意,因此提高长术语的发现能力很有意义。为此可增加候选术语长度在术语度值中的权重,因此本文认为用代替更为妥当。

综上所述,IC – value 计算公式如下:

$$\text{IC-value}(a) = \begin{cases} |a| \cdot f(a) \cdot \log\left(\frac{N}{g(a)}\right) & \text{当 a 无嵌套} \\ |a| \cdot \left(f(a) - \sum_{b\in T_a} f(b)\right) \cdot \log\left(\frac{N}{g(a)}\right) & \text{其他情况} \end{cases} \quad (1)$$

其中,a 表示候选术语,IC – value(a) 指候选术语 a 的 IC – value 值,|a| 表示候选术语 a 的长度,即候选术语包含的字数,f(x) 表示 x(x 取 a 或 b) 在文档集中出现的频次,g(a) 表示候选术语 a 的文档频率,b 是候选术语 a 的嵌套候选术语,T_a 表示候选术语 a 的嵌套候选术语集合。需要注意的是,在计算 f(b) 时较短的嵌套候选术语的频次要扣除嵌套它的候选术语的频次。举例来说,“细胞”的嵌套词语有“细胞免疫”、“细胞免疫应答”、“细胞免疫功能”等。在计算细胞的嵌套候选术语的总频次 $\sum_{b\in T_a} f(b)$ 时,“细胞免疫”的 f(b) 值应为“细胞免疫”的词频减去“细胞免疫应答”和“细胞免疫功能”的词频。

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

根据公式(1)计算得到每个候选术语的 IC - value 值,其值越高,成为术语的可能性就越大。

4　实验及结果分析

4.1　实验及其结果

本文选取 2007 年到 2011 年共 1 000 篇有关乙型肝炎相关论文的摘要,并以此进行术语抽取实验。

(1)利用停用词对摘要进行预处理,并采用基于串频统计的候选术语抽取算法从预处理文本中抽取候选术语,将阈值设置为 2,并且抽取长度大于 1 的候选术语,共得到 3 916 个候选术语。

(2)对抽取的候选术语集进行分词及词性标注之后,使用非术语语言学规则进行过滤,将明显不能作为术语的候选术语过滤掉。同时发现候选术语集中包含一些地名,本文利用候选术语是否包含"省"、"市"、"县"将地名过滤。经过以上处理,最终得到 3 270 个候选术语。

(3)对于最终得到的 3 270 个候选术语,计算其 TF - IDF、C - value 和 IC - value 值,同时为了与文献[15]的方法比较,也计算 V - value 值(由于在此仅用文献摘要,因此在计算 V - value 时没有使用加权 TF - IDF)。然后分别按照 TF - IDF、C - value、V - value 和 IC - value 值大小进行排序,并按照排序结果选取 Top500。人工判定 3 270 个候选术语是否为术语,并以此来判定 Top500 中的术语是否正确,从而计算出不同方法的 Top500 术语选择的正确率。同时,从 1 000 篇摘要中随机抽取 50 篇摘要,人工识别出其中的术语,共计 265 个,来检验不同方法 Top500 的召回率。其中,准确率和召回率的计算公式如下:

$$准确率=\frac{正确的术语数}{抽取出的术语总数}\times 100\% \quad (2)$$

$$召回率=\frac{正确的术语数}{语料中包含的术语数}\times 100\% \quad (3)$$

根据公式(2)与公式(3),得到实验结果的准确率与召回率分别如表 2 和表 3 所示:

表 2　TF - IDF、C - value、V - value 和 IC - value 方法 Top500 术语选择的准确率

方法	准确率				
	Top100	Top200	Top300	Top400	Top500
TF - IDF	57.00%	60.00%	63.00%	62.00%	64.80%
C - value	68.00%	71.50%	72.00%	73.25%	73.60%
V - value	57.00%	63.50%	68.00%	67.50%	69.00%
IC - value	80.00%	80.50%	78.00%	78.50%	77.80%

表 3　TF - IDF、C - value、V - value 和 IC - value 方法 Top500 术语选择的召回率

方法	召回率				
	Top100	Top200	Top300	Top400	Top500
TF - IDF	9.81%	12.83%	17.74%	20.38%	23.40%
C - value	11.70%	16.60%	20.38%	23.40%	25.28%
V - value	10.57%	15.09%	20.75%	24.15%	27.55%
IC - value	10.94%	19.62%	23.02%	27.55%	29.81%

4.2　实验结果分析

由表 2 可以看出,IC - value 方法的 Top500 术语抽取的准确率要比其他三种方法高。在抽取的 Top100 术语中,IC - value 方法要比 TF - IDF 和 V - value 方法高出 23%,比 C - value 方法高出 12%。基于 C - value 改进的 V - value 方法抽取术语的准确率远不如 IC - value 方法,甚至不如 C - value 方法,但比 TF - IDF 方法略好。由此可见,IC - value 方法有效提高了术语抽取的准确率,效果优于 V - value 方法。

由表 3 可以看出,针对随机选取的 50 篇摘要中的 265 个术语,IC - value 方法的 Top500 术语抽取召回率比 C - value、TF - IDF 和 V - value 效果略好,但是,4 种方法 Top500 的召回率都在 30% 以下,召回率普遍较低。分析 50 篇摘要中的 265 个术语,发现其中共有 62 个术语在整个文档集中的出现频次为 1,本文方法无法抽取。并且 265 个术语中在文档集出现频次少于等于 10 的术语共 148 个,而其中被 TF - IDF、C - value、V - value 和 IC - value 方法识别的术语都不足 10 个,分别为 0、8、4 和 8,说明 4 种方法对低频术语的识别能力较差。由此可见召回率偏低的主要原因是低频术语较难抽取。

为了更好地评价抽取结果,对 4 种方法的 Top500 进行更为深入的分析。在对 Top500 中正确抽取的术语分析后发现,IC - value、C - value、V - value 和 TF - IDF 方法得到的长度大于 2 的术语分别为 335、320、254 和 196 个。可见 IC - value 发现长术语的能力比 C - value 方法稍优,远优于 V - value 和 TF - IDF 方法。

另外,在由 TF - IDF 和 C - value 方法得到的 Top500 中有许多没有意义的破碎子串,比如"慢性乙"、"夫定"、"荧光定"、"慢性肝"等。统计 4 种方法 Top500 中破碎子串的个数如表 4 所示。

可以发现 IC - value 方法能有效地区分破碎子串和术语。主要原因是在 TF - IDF、C - value 和 V - value 的

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

表4　TF－IDF、C－value、V－value 和 IC－value 方法 Top500 中破碎子串的数量

	TF－IDF	C－value	V－value	IC－value
破碎子串数	36	21	41	1

计算公式中没有考虑到破碎子串的问题，在计算相应值时不按其破碎子串独立出现的次数计算，导致其对应值较高。而 IC－value 方法使用 $f(a) - \sum_{b \in T_a} f(b)$（子串独立出现的频次）计算术语度值。

5　结　语

由于语言学规则很难全面总结，本文使用一种基于串频统计的候选术语抽取算法提取候选术语。然后采取逆向的方法，利用候选术语的词性将一些明显不能作为术语的候选术语过滤掉。针对术语选择问题，考虑逆文档频率、破碎子串和候选术语长度的权重三个方面，对 C－value 方法进行了改进。最后，以 1 000 篇乙型肝炎相关论文摘要对本文提出的方法进行实证研究。结果证明，本文提出的 IC－value 方法在准确率和召回率方面都要优于 TF－IDF 、C－value 以及 V－value方法，且有较强的长术语发现能力，识别公共破碎子串的效果十分明显。

参考文献:

[1] 冯志伟. 现代术语学引论[M]. 北京: 语文出版社，1997.（Feng Zhiwei. An Introduction to Modern Terminology [M]. Beijing: Language & Culture Press，1997.）

[2] 王强军，李芸，张普. 信息技术领域术语提取的初步研究[J]. 术语标准化与信息技术，2003(1)：32－34.（Wang Qiangjun，Li Yun，Zhang Pu. Automatic Term Extraction in the Field of Information Technology [J]. *Terminology Standardization and Information Technology*，2003(1)：32－34.）

[3] 安纪霞，李锡祚，宋冰，等. 服务于词典编纂的特定领域专业术语自动抽取[J]. 计算机与数字工程，2007，35(11)：53－56.（An Jixia，Li Xizuo，Song Bing，et al. Service in Dictionary Compilation of Specific Areas of Professional Term Automatic Extraction [J]. *Computer and Digital Engineering*，2007，35(11)：53－56.）

[4] Foo J，Merkel M. Using Machine Learning to Perform Automatic Term Recognition [C]. In: *Proceedings of the LREC* 2010 *Workshop on Methods for Automatic Acquisition of Language Resources and Their Evaluation Methods*，Valletta. 2010：49－54.

[5] Krauthammer M，Nenadic G. Term Identification in the Biomedical Literature [J]. *Journal of Biomedical Informatics*，2004，37(6)：512－526.

[6] Kageura K，Umino B. Methods of Automatic Term Recognition：A Review [J]. *Terminology*，1996，3(2)：259－289.

[7] 潘虹，徐朝军. LCS 算法在术语抽取中的应用研究[J]. 情报学报，2010，29(5)：853－857.（Pan Hong，Xu Chaojun. Application of LCS－based Algorithm in Chinese Term Extraction [J]. *Journal of the China Society for Scientific and Technical Information*，2010，29(5)：853－857.）

[8] Damerau F J. Generating and Evaluating Domain－oriented Multi－word Terms from Texts [J]. *Information Processing & Management*，1993，29(4)：433－447.

[9] 张锋，许云，侯艳，等. 基于互信息的中文术语抽取系统[J]. 计算机应用研究，2005，22(5)：72－74.（Zhang Feng，Xu Yun，Hou Yan，et al. Chinese Term Extraction System Based on Mutual Information [J]. *Application Research of Computers*，2005，22(5)：72－74.）

[10] Gelbukh A，Sidorov G，Lavin－Villa E，et al. Automatic Term Extraction Using Log－Likelihood Based Comparison with General Reference Corpus [C]. In: *Proceedings of the Natural Language Processing and Information Systems，and the* 15*th International Conference on Applications of Natural Language to Information Systems*. Berlin，Heidelberg: Springer－Verlag，2010：248－255.

[11] 周浪，史树敏，冯冲，等. 基于多策略融合的中文术语抽取方法[J]. 情报学报，2010，29(3)：460－467.（Zhou Lang，Shi Shumin，Feng Chong，et al. A Chinese Term Extraction System Based on Multi－Strategies Integration [J]. *Journal of the China Society for Scientific and Technical Information*，2010，29(3)：460－467.）

[12] 岑咏华，韩哲，季培培. 基于隐马尔科夫模型的中文术语识别研究[J]. 现代图书情报技术，2008(12)：54－58.（Cen Yonghua，Han Zhe，Ji Peipei. Chinese Term Recognition Based on Hidden Markov Model [J]. *New Technology of Library and Information Service*，2008(12)：54－58.）

[13] Frantzi K，Ananiadou S，Mima H. Automatic Recognition of Multi－word Terms：The C－value/NC－value Method [J]. *International Journal on Digital Libraries*，2000，3(2)：115－130.

[14] 中英文混合停用词表[EB/OL]. [2012－11－20]. http://www.smartpeer.net/myfiles/stopwords－utf8.txt.（A Mixture of English and Chinese Stoplist [EB/OL]. [2012－11－20]. http://www.smartpeer.net/myfiles/stopwords－utf8.txt.）

[15] 许德山，张智雄，王峰，等. 上下文分析与统计特征相结合的英文术语抽取研究[J]. 现代图书情报技术，2010(12)：28－32.（Xu Deshan，Zhang Zhixiong，Wang Feng，et al. English Term Extraction Based on Context Analysis & Statistical Characteristic [J]. *New Technology of Library and Information Service*，2010

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

(12):28 - 32.)

[16] 李超,王会珍,朱慕华,等. 基于领域类别信息 C - value 的多词串自动抽取[J]. 中文信息学报,2010,24(1):94 - 98. (Li Chao, Wang Huizhen, Zhu Muhua, et al. Exploiting Domain Interdependence for Multi - Word Terms Extraction[J]. *Journal of Chinese Information Processing*, 2010, 24(1):94 - 98.)

[17] 韩红旗,朱东华,汪雪锋. 专利技术术语的抽取方法[J]. 情报学报,2011,30(12): 1280 - 1284. (Han Hongqi, Zhu Donghua, Wang Xuefeng. Technical Term Extraction Method for Patent Document[J]. *Journal of the China Society for Scientific and Technical Information*, 2011, 30(12): 1280 - 1284.)

(作者 E - mail: huap2011@istic.ac.cn)

GigaOM 预测 2013 年云计算领域的 5 大变革

云已经从概念变成了现实。2013 年将看到更多公司提供企业级的 IaaS 应用程序和更可行的混合云模型,使云计算在 2013 年爆发式增长。

(1) 公共云可处理企业应用

有消息称,已经有一些财富 1 000 强公司在亚马逊的公共云上进行测试和开发,甚至纳斯达克(NASDAQ)也是 AWS(Amazon Web Service)的一个客户。然而,在关键任务的应用程序方面受到严格监管的金融和医疗行业,其中许多公司不会在公共云中保存任何数据或应用程序。比如一些银行甚至不允许员工使用 AWS,更不用说进行部署。

这对于亚马逊(和微软 Azure)来说是一个巨大的障碍。AWS 去年和 Eucalyptus 达成协议,使企业能够将 Eucalyptus 私有云和 AWS 组成一个混合模型。像 CloudVelocity 这样的创业公司声称他们可以"克隆"内部的工作负载到 AWS 上,并提供完整的安全保障。这是一个很大的承诺,2013 年我们将看到更多的这样的声明。

与此同时,企业软件巨头 VMware 和微软也必须向他们原有的客户证明他们的云技术是符合标准的,并且具有新的前景。

(2) 不成则败的惠普

在过去的几年里,硬件厂商——戴尔、EMC、惠普和 IBM 都争相证明其与云计算这个新世界的关联性。

惠普在即将到来的一年中将最热闹,首席执行官 Meg Whitman 恳求投资者等待惠普的一个"多年的转变"。惠普多年的管理混乱以及最近收购 Autonomy 的可疑举动,已将其推到十字路口。这项 111 亿美元的收购旨在构建惠普在大数据和云计算方面的信誉。可以肯定,它并没有达到预期的效果。

现在,惠普推出了其基于 OpenStack 的计算云。惠普如果想成功翻身,它的企业客户基础必须巨大,并且他们有足够的信心使用惠普的云服务,否则终将失败。

(3) OpenStack 将大放光彩

现在,Rackspace 已经不再将自己定位在 OpenStack"父亲"的角色上,OpenStack 已经变成由多个厂商共同管理,OpenStack 能否与 AWS 在公共云端上进行竞争,能否与 CloudStack、Eucalyptus 和 OpenNebula 等其他开源云之间进行竞争,让我们拭目以待。

2012 年对于 OpenStack 来说是丰收的一年,惠普、Internap、红帽和 Rackspace 都推出了基于 OpenStack 的云服务。Nebula 的 OpenStack 设备全面上市的日期越来越近,还有其他的云选择也不断出现,如亚马逊和谷歌 API 兼容的私有云。更多的公司正在围绕 OpenStack 构建服务,例如 Mirantis 刚刚推出了自己的自助 OpenStack 服务。

(4) 基础设施不再局限于数据中心

早在 2008 年,谷歌就推出了"数据中心是计算机"的想法,随着其推出能横跨 5 个数据中心进行内容同步的 Spanner 数据库,我们已经进入一个新的基础设施领域。现在的数据中心还不是一台计算机,而是连接在网络上的一堆设备。不只是谷歌有这种想法。Facebook 也进行了类似的努力,向运营商租赁光纤,以扩展其基础设施的覆盖范围。2013 年,我们将看到更多的数据中心运营商进行交易,他们将不得不扩展他们数据中心之外的基础设施,探索如何在大型分布式网络中保持同步。

(5) 软件定义的一切并没有变得更容易

"软件定义网络"是 2012 年的大热门词汇,还出现了"软件定义存储"和"软件定义的数据中心"等词汇。这些词汇的基本思想是把虚拟化给计算带来的灵活性同样赋予网络、存储和数据中心。但是,就像任何新的领域都可能会破坏一些厂商的既得利益一样,很多营销人员不看好软件定义的网络。尽管我们期望看到很多生产部署网络虚拟化的表现,但目前还没有看到多大的进展。

(编译自: http://gigaom.com/cloud/what - well - see - in - 2013 - in - cloud - computing/)

(本刊讯)

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net