Probase：一个概率分类文本理解

文涛武 [1] *　　红松李 [2]　　海迅王 [2]　　肯尼Q.朱 [3] *

[1] 威斯康星州，威斯康星州麦迪逊大学，美国
[2] 微软亚洲研究院，北京，中国
[3] 上海交通大学，上海，中国

wentaowu@cs.wisc.edu ，{hongsli，haixunw} @ microsoft.com，kzhu@cs.sjtu.edu.cn

抽象

知识是必不可少的理解。正在进行的Infor公司，mation爆炸强调必须使机器bet-之三理解人类语言的电子文本。许多工作一直致力于为此目的建立通用的本体或分类。然而，没有任何现有的本体是否有“普遍理解”所需要的深度和广度。在本文中，我们提出了一个通用的，概率分类是比较compre-角度，系统化的比任何现有的。它包含了270万点的概念从1.68十亿网页的语料库自动利用。不像对待知识黑白传统的分类法，它使用概率来它包含不一致的，模糊的，不确定的信息模型。我们目前的分类标准是如何构建的，其概率模型的细节，

分类和主题描述

H.3.4 [ ]信息存储和检索： 系统和软洁具

一般条款

算法，设计，试验

关键词

知识库，分类，语篇理解

1。 介绍

网络已经成为世界上最大的数据仓库之一。但是Web上的数据大多是在自然语言中，这是不好的机器访问结构化和DIF网络邪教文本。要解锁的信息宝库，我们必须使机器自动亲塞斯网络数据。换句话说，机器需要这不同于derstand文本的自然语言。

一个重要的问题是，什么是词“ *了解* ”这里的意思？请看下面的例子。对于人类来说，当

***这项工作是在微软亚洲研究院完成。肯尼问：朱部分被国家自然科学基金资助610330**02和61100050的支持。

SIGMOD'12， **2012年5月20-24日，亚利桑那州斯科茨代尔，美国。版权所有2012 A**
CM 978-1-4503-1247-9 / 12月5日... $ 10,00。

我们看到“1881年10月25日”，我们认识到它作为一个日期，但我们大多数人不知道它是什么。但是，如果我们给予多一点背景，说的日期嵌入在下面这段简短的文字的“毕加索，1881年10月25日，西班牙”，我们大多数人已经猜到（正确地）的日期代表巴勃罗·毕加索的生日。我们有能力，因为我们具有某些knowl-边缘要做到这一点，在这种情况下，**“ *与人相关联的最重要的日子之一是他的生日* ”。再举一个例子，*考虑*包含以下两句话“ *如* ”这样一句话：“ *House - 临近拉宠物* 以外 *小狗* 如 *猫* ... ”和“ *家庭宠物* 以外 *动物* 如 *爬行动物...* ”。人类不觉得自己很暧昧。潜意识中，他们分析以不同的方式，以获得正确的**语义的句子：“家庭宠物如猫”的第一个句子，“动物如爬行动物”为第二，而机器无法理解为什么“狗如猫”和“家庭宠物如爬行动物”是不可能的解释。究其原因是因为人类有背景知识。

事实证明，有什么需要一个人来了解上面的两个例子无非是对知识更 *概念（* 例如，人，动物等）的能力， *概念化（* 例如，猫是动物）。这绝不是巧合。心理学家格雷戈里Mur- PHY开始了他很高的评价书的声明“ *概念是共同拥有我们的心理世界胶水* ”[22]。自然杂志书评中指出“ *没有概念，就没有心理世界的第一个地方* ”[4]。

人们用分类法和本体代表和组织观念。在开域理解文本（例如，在网络上understan d- ING文本）是非常具有挑战性的。多样性和人类语言的复杂性需要分类/本体成Cap-在**各种粒度TURE概念 *一切* 域。在SPECI音响C结构域，尽管许多分类法/存在本体，只有**通用那些少数的（表1）是可用的。这些税 - onomies /本体共用两个关键的限制，这使得它们在通用的理解不太有效。

首先，现有的分类有 *有限的概念空间。* 最税 - onomies由称为手动过程构造 *策展。*

这种费力，费时且昂贵的过程限制的范围，从而建立了分类法的规模。例如，项目的C yc [18]，被众多领域知识专家25年继续努力的，包含约120,000概念。为了克服这一瓶颈，一些开放领域knowledgebases，例如，游离碱[5]，依靠社会各界的努力，以增加规**模。如何 - 以往，而他们有几个SPECI科幻ç概念几乎完全覆盖（例如， *书籍，音乐* 和 *电影）， 他们缺乏等诸多概念的一般杯盖 - 年龄。最近，自动分类结构的方法，如**Kno**wItAll [12]，TextRunner [2]，YAGO [35]，和NELL [7]，已在焦点，但它们仍然具有在概念空间方面具有有限的范围和覆盖面。

有限的概念空间限制的理解在 *胆* 水平。请看下面的句子：

Ë XAMPLE 1。 *"我们如何与大公司在中国，印度，巴西等国竞争呢？"*

哪些国家是中国，印度和巴西？什么是最大的公司呢？大多数现有的分类标准包含CON组CEPT *公司*， 并有中国，印度和巴西在概念

国家。 然而，这些概念太笼统，不利于理解。为了揭开语义编码这里，机器需要的很多知识 *科幻NER* 的概念，如 *在中国最大的compa-新兴工业化经济体*， *发展中国家*， 和 *金砖四国*。 Unfor- tunately，没有一个现有的分类标准中包含了这些概念。

理解精细程度也是许多其他的IM portant任务，如可取 *命名实体识别（* NER）[23]和

词义消歧（ WSD）[24]。NER寻找定位，并与文中提到的标签现实世界的实体 *类型（* 即，概念）。虽然在NER早期研究是CON音响斯内德 *粗粒* 命名实体类，如 *人* 和 *地点*， 它erally gen-同意 *科幻细粒度* NER [14，15]（即，通过使用更多的SPECI音响Ç子类别）为广泛的网络应用中，包括信息检索（IR），信息EX-牵引（IE），或查询的应答（QA）更多好处音响官方。WSD支配识别所述文本中使用的词的意义（即，意思）的过程。知识信息源，例如分类法和本体是WSD，其通过其类别（即，概念）区分词语的感官心理手达部件。它有不同的NLP应用程序，以最好地利用词义的区别[33]需要不同的粒度感得到了广泛的观察。所有这些拖至应用程序所需要的分类/本体包含一个 *丰富* 设定的概念，用各种粒度。

现有的分类法概念的数	
游离碱[5]	1450
WordNet的[13]	25229
WikiTaxonomy [26]	111654
YAGO [35]	352297
DBpedia中[1]	259
ResearchCyc [18]	≈120000
KnowItAll [12]	N / A
TextRunner [2]	N / A
OMCS [31]	N / A
NELL [7]	123
Probase	2653872

表1：开放域分类规模

其次，现有的分类处理知识 *黑与白*。
他们认为，一个知识库应该提供标准，良好德科幻奈德和一致可重复使用的信息，因此所有CON组cepts和关系包括在这些分类都保持"reli- giously"干净。然而，许多现实世界的概念，如 *大公司*， *最好的大学* 和 *美丽的城市* 没有 *德无服*

边界，并在本质上是 *模糊*。 许多这些含糊CON组cepts都在网络细粒度的理解是有用的。举例来说，有在世界大学的几万，而只有hun-上将小于它们被认为是 *最好*。 此外，自动税 - onomy建设进程，同时降低成本和提高生产力，是永远不会完美。他们可能会导致错误和IN一致性成这样建立的分类法。如果不能含糊和矛盾建模一个很好的方式，所有现有的分类选择要么排除模糊的概念或忽略的不一致。

在本文中，我们认为，容纳和建模在分类学这种不确定性可以在概念非常有用的。看到这一点，让我们在例1中的句子根据不同的应用更深入的了解，我们可能需要了解：1）什么是" *大公司* " 意思？和ii）什么是" *中国，印度，巴西* "意味着什么？

显然还有数以百万计的公司在这些国家，但其中只有极少数被认为是 *最大的公司*。由于术语" *最大* "是 *主观*， 要了解什么是 *最大compa-新兴工业化经济体*， 我们具体化这个模糊的概念一组最 *典型*

实例，如中国移动，塔塔集团，巴西国家石油公司和。在另一方面，这三个实例 *中国*，*印度*， 和 *巴西*，
可以解释为 *国家，大国，发展中国家，金砖四国*， 要么 *新兴市场*。 所有这些选择都心病 - 矩形，但作为一个群体一起， *金砖四国* 和 *新兴市场*

可能是最好的抽象，因为他们是最 *典型*
与"紧"的概念来描述的三个实例。有了这个泛化，人们甚至可以提出一个第四个实例，*俄国*，
完成句子。
从上面的例子中，我们可以看到，概念化可以在两个方向着手：

- ***实例：*** 给出一个概念，推断其典型和可能的情况下（例如，从 *大公司* 至 *中国移动*，*塔塔集团*， 等等）。
- ***抽象：*** 给定的一个或多个实例[1]，推断它们属于如（典型和可能的概念，从 *中国*，*印度*，*巴西* 至 *新兴市场* 要么 *金砖四国*）；

在任一方向，不确定性是固有的，概率起到推理的重要作用。

在本文中，我们介绍Probase，通用，通用，概率分类自动从语料库构建

1.6十亿网页。该Probase分类是在三个方面独特之处：

1.它是由一种新的框架，它由一个iter- ative学习算法来提取建 *ISA* 从网络文本对（见第2），和一个分类构造算法给这些成对的连接成一个分级结构（见

3）。得到的分类具有最高的精度（92.8％），规模最大的自动化网络的规模分类推理研究的报道。

2.这是第一个通用分类标准是需要tomodel它拥有的知识proba- bilistic方法。在Probase Knowl-边缘不再是黑色和白色。每个事实或关系与一些概率来衡量其合理性和典型性相关。可信度是detect- ING错误和集成异构知识来源是有用的，而典型性是概念化和推理有用。这样的概率处理使得Probase更好地捕获到的波形人类语言（见第4节）的语义。重新分的工作[34，39，37]基于在Probase所知，这种概率治疗演示此框架的有效性，这将进一步在部分讨论

5.3。

3.它是全球最大的通用分类标准在网络上充分automati-凯莉构建fromHTML文本。Probase有将近270万的概念，比YAGO大8倍一个巨大的概念空间，这使得它的概念空间（表1）而言为最大taxon- OMY。除了流行的概念，如"城市"和"音乐家"，这已经在几乎所有的通用分类标准，它也有几万SPECI科幻ç概念，如"可再生能源技术"，"气象现象"和"共同

[1] Probase还支持从实例中，AT-悼念和动作的混合物抽象。例如，来自推断 *总部，AP-PLE* 至 *公司*， 或 *德国入侵波兰* 至 *战争*。 如何 - 以往，在本文中，我们侧重于概念和实例。

睡眠障碍”，它不能在游离碱，的Cyc，或任何其他分类中找到（见第5章）。

关于Probase更多信息，包括启用Probase-拖至应用程序[34，39，37]，和Probase taxon- OMY小摘录，可以被发现在 http://research.microsoft.com/ probase /。 目前，我们正在努力使Probase taxon- OMY提供给公众。

2。 迭代提取

我们提出了一种新的迭代学习框架，旨在以高精确度和高召回AC- quiring知识。知识获取由两个阶段组成：1）信息提取，以及ii）数据清理和集成。大量的工作已经在数据清洗和整合[17，19，20] Probase已经完成。在本文中，我们着重于第一个阶段：信息提取。

信息提取是一个迭代过程。大多数现有的AP- proaches引导上句法模式，即，用于随后的提取每次迭代音响NDS更句法模式。我们的方法，在另一方面，白手起家直接的知识，也就是我们利用现有的知识理解课文和掌握更多的知识。在下文中，我们描述的最先进的国有工作的局限性2.1节中，存在的问题，并在第2.2节的挑战，并在第2.3节我们新的迭代学习框架。

2.1句法语义与迭代

状态的最先进的信息extractionmethods，包括知道 - ItAll [12]，TextRunner [2]，和NELL [7]，依赖于一个迭代（boot-捆扎）的方法。它开始于一个组的种子的实例和/或种子的图案。从例子中，它获得的新模式，网络连接t时的例子。然后，它使用新的模式，从数据中提取更多的EX- amples。迭代过程是在语法层面。它具有预防深知识获取的限制。我们的目标是打破这一障碍，并在语义，或知识水平进行萃取。

句法迭代。 **高品质的句法模式valu-能够信息提取。假设我们感兴趣的是网络的ND荷兰国际集团 *ISA* 关系（在知识库中的最重要的关系）。我们可以用赫斯特模式[16]启动。**

ID	图案
1	***NP* 如 { *NP* , *}*((或者| 和) } *NP***
2	**这样 *NP* 为{ *NP* , *}*((或者| 和) } *NP***
3	***NP* { , *}包含 { *NP* , *}*((或者| 和) } *NP***
4	***NP* { , *NP}*{ , *}和别的 *NP***
五	***NP* { , *NP}*{ , *}或其他 *NP***
6	***NP* { , *}特别是{ *NP* , *}*((或者| 和) } *NP***

表2：赫斯特图案（ *NP* 代表 *名词短语* ）

使用上述赫斯特模式，我们可以从文字中获得知识。例如，给定一个句子，“.....家养动物如猫.....”，我们得到的关系：“猫 *ISA* 动物”。句法迭代的想法是，为了到FI ND多种关系，我们需要更多的句法模式。因此，我们努力探索当前之间存在更多的句法模式 *ISA* 对（其在 - CLUDE“猫 *ISA* 动物”），并利用它们来获得更多的 *ISA* 对。

这个过程已经通过了最多的信息提取方法。然而，只专注于语法，并拥有这些lim- it ations：

- *句法模式已经限制抽取功率。* **自然语言模糊，句法模式是孤独**

没有强大到足以应对这种不确定性。考虑了一句“狗比其他动物...如猫.....”。Kn owItAll将提取物（猫 *ISA* 狗），而不是（猫 *ISA* AN-进制）。因为句法结构具有内在ambigu-的OU（如我们在第1节中所述），重新连接宁语法规则不会在这种情况下帮助。

- *高品质的句法模式是罕见的。* **从引导获得的语法PAT-燕鸥经常有低质量的。举例来说，假设我们要到FI国家的第二实例，即（X *ISA* 国家）。从一组种子的国家，我们可以得到句法模式，诸如“用X战”，“x的inva-锡永”和“x的占领”。但是，从这样的模式，我们可以推导出错误的情况下，例如，“X =地球。”这个问题被称为语义漂移[7]。为了解决这个问题，创建复杂的“鉴别”，除去低质量的句法模式。不幸的是， *ISA* 关系，其余的模式大多是赫斯特的模式。这样的中心思想“更句法模式能产生更多的成果”是不是真的有效。**

- *召回是土木工程智精密的牺牲网络。* **由于自然语言模糊，句法模式具有低品质，最先进的国有方法有SACRI科幻CE召回的精度。例如，当提取 *ISA* 对，他们专注于在 - 立场是 *专有名词。* 因此，它不能推导出knowl-边缘（猫 *ISA* 动物）从简单的句子，如“动物如猫”。但是，这种 *ISA* 关系在形成不同的知识分类是必不可少的。此外，大多数的方法也限制的概念是名词，而不是一个名词短语。例如，从一个句子“.....工业化，如美国和德国等国家.....”，只有（美国 *ISA* 国家）是EX-牙牙，而不是（美国 *ISA* 工业化国家）。**

语义迭代。 Probase执行的迭代学习
知识 或语义水平。鉴于这句话，“狗比猫等国内其他ani- MALS”，Probase意识到，有两种可能的读数：（猫 *ISA* 狗和猫 *ISA* 家畜）。如果Probase一无所知猫，狗，家畜（这是在第一个迭代的情况下），它不能决定哪一个是更可能的。因此，该句子被丢弃。然而，第二次迭代开始时，Probase已经已经获得了很多知识，也知道（家养动物 *ISA* 动物），和（猫的频率 *ISA* 动物）比（猫要高得多 *ISA* 狗）。当这种差异是高于某一阈值，就可以正确地在两个可能的读数之间进行选择。

换句话说，在Probase，获得新knowl-边缘的力量并非来自使用更句法模式。事实上，一个固定设置句法模式，即，赫斯特模式，在每次迭代中使用的。相反，权力来自现有knowl-边缘。

在命名实体识别（NER）一些以前的工作也解决类似的问题。唐尼等。[10]提出了后续ING问题：我们怎么知道“...公司，如宝洁.....”在谈论两家公司{ *宝洁，宝洁}* 或单一的公司，其名称为 *宝洁？* 他们的想法是使用逐点互信息（PMI）来衡量的通货膨胀associ- *普罗克特* 和 *赌。* 这类似于在第2.3.3节中讨论我们的方法，我们注意到，频率

宝洁 作为一个术语是比所述频率高得多 *普罗克特* 单独出现。与他们的工作相比，我们的AP-proach较为一般，因为我们不局限于PMI。其实，我们可以采取的任何现有的知识，我们已经了解到的优势。举例来说，如果我们知道我们在谈论“卡通”（超级概念，这是一个上下文 *ISA* 提取），它

更有可能的是 *汤姆和杰瑞* 应作为一个单一的IN-姿态被处理，而不是两个实例{ *汤姆，杰里*}。

2.2问题德网络nition

在本文中，我们提出我们的迭代学习框架中提取的设置 *ISA* 从Web文档对。不像先进依赖于发现更多的句法模式获得新的知识信息提取方法国有，我们的它 - 关合作使用一个固定设置句法模式（赫斯特模式），并依赖于使用现有的知识，了解更多的文字，并获得更多的知识。

从任何匹配赫斯特模式的一句话，我们要获得

$$S = \{(X, Y_1), (X, Y_2), \ldots, (X, Y_n)\}$$

哪里 X 是上位概念（或超概念），和{ $\bar{y}_1$，
· · · ，$\bar{y}_N$} 是其下属的概念（或子概念）。例如导出，从句子

"……在热带国家 如 *新加坡，马来西亚，* …"[2]

我们得出 $S = \{($ 热带国家，新加坡），（热带国家，马来西亚）}。

自然语言充斥着模糊性，和句法PAT-燕鸥不能单独用模糊处理。下面是一些应试普莱斯（在我们的语料库中找到）：

ËXAMPLE 2。 *例句。*

1）... 动物比其他狗 如 *猫* ...

2）... 经典电影 如 *随风而逝* ...

3）... 公司 如 *IBM，诺基亚，宝洁*.....

4）... 在北美，欧洲，中东的代表，
澳大利亚，墨西哥，巴西，日本，中国， 和别的 国家.....如果我们只依靠句法模式，1）*小狗* 将被提取的超概念，而不是 *动物;* 2）因为没有被提取 *随风而逝* 是不是一个名词短语; 3）*普罗克特* 和 *赌* 被视为两家公司; 4）*北美，欧洲，* 和 *中东地区* 被错误地提取作为国家。

2.3框架

我们的框架关注 *理解。* 在许多情况下，SE-mantics需要补充正确提取的语法。随着迭代的进步，我们获得了更多的知识。使用知识，我们可以有一个更好的理解语义，这更增加了我们的提取框架的力量。

具体来说，我们提出了一个反复的学习过程。在每一轮信息提取的，我们积累了，我们有高置信度是正确的知识。然后，我们用这些知识在下一回合中，以帮助我们提取我们以前遗漏的信息。我们执行此过程 *迭代* 直到可以提取没有更多的信息。

更具体来说，让 Γ 表示我们目前拥有的知识，
即一套 *ISA* 我们已经发现了对。对于每一个（X, Y）$\in$
Γ，我们也保持计数 $N(X, Y)$，这表明有多少次
（X, Y）发现。原来，Γ 是空的。我们寻找 *ISA* 对在文中，我们使用 Γ 以帮助识别其中有效问卷。我们扩大 Γ 通过将新发现的对，这进一步模式，提高了我们的力量，以确定更有效的对。表3概括整篇文章中使用的重要符号。

[2] 带下划线的术语是超级的概念，而斜体方面是它的子概念。

符号含义	
Γ	集 *ISA* 从所述语料库中提取对
(X, Y)	一个 *ISA* 对具有超概念 X 和子概念 $\bar{y}$
$N(X, Y)$	次 #（X, Y）在语料库中发现
$_{小号}$	一个匹配任何赫斯特模式的句子
$X_{小号}$	句子的候选超概念 S y时 $_{小号}$
	句子的备选子概念 $SX_{-世}$
	概念 X 与意义 $/T_{九}$
	本地分类与根 $X_{-世}$
$P(X, Y)$	的合理性 *ISA* 对（X, Y）
$T(我|X)$	实例的典型性 *一世* 给出的概念 X
$T(X|I)$	这一概念的典型性 X 给出的实例 *一世*

表3：不缩

算法1： *ISA* 萃取

输入：S，从网页语料匹配赫斯特指出，判刑
模式

输出：Γ，设置 *ISA* 对

1 $\Gamma \leftarrow \emptyset$;
2重复3
　的foreach $_{小号} \in _{小号}$ 做4
　　$X_S, \bar{y}_{小号} \leftarrow$ *SyntacticExtraction*（一个或多个）;
五　　如果| $X_{S|>}$ 1 然后6
　　　$X_{小号} \leftarrow$ *SuperConceptDetection*（$X_S, \bar{y}_S, \Gamma$）;
7　　8月底
　　如果| $X_{S|=}$ 1 然后9
　　　$\bar{y}_{小号} \leftarrow$ *SubConceptDetection*（$X_S, \bar{y}_S, \Gamma$）;
10　　添加有效的 *ISA* 对来 Γ;
11　　端12
　结束13至 *没有对新的加* Γ;

14回 Γ;

算法1概述了我们在较高水平的方法。它反复扫描的句子集，直到没有更多的对可以identi网络版的。程序 *SyntacticExtraction* 网络连接NDS *候选人* 超级概念 $X_{小号}$

和 *候选人* 子概念 $\bar{y}_{小号}$ 从句子 *秒。* 如果存在一个以上的候选人的超级概念，我们称之为程序 *SuperConcept-检测* 减少 $X_{小号}$ 到一个单一的元素。然后，程序 *子ConceptDetection* 滤池出不可能的子概念 $\bar{y}_{秒}$。最后，我们添加新发现 *ISA* 对到结果。由于新的结果，我们也许能够识别更多的对，所以我们再次扫描句子。我们将在下面详细的三个过程。

2.3.1句法提取

程序 *SyntacticExtraction* 检测候选超概念
$X_{小号}$ 和子概念 $\bar{y}_{小号}$ 在一个句子 *秒。*

如实施例2中，名词短语的句子1中示出），该 *最近的* 该模式的关键字可能 *不* 是正确的超级概念。因此，$X_{小号}$ 应该包含所有可能的名词短语。对于句子1），我们确定候选人的超级概念 $X_{1)=}$

{动物，狗}。 正如一些以前的工作[12，27]，我们fur-疗法需要在每一个元素 $X_{小号}$ 必须是一个名词短语
复数形式。 其结果是，对于句子"……日本以外的国家 如 *美国* …"，该组候选超级概念包含了'国家'，而不是'日本'。

这是比较难以确定 $\bar{y}_{秒}$。首先，如在森唐塞2）中所示，子概念可 *不* 是名词短语。第二，如在句子3中所示），分隔符，如" *和* "和" *要么* "本身可能出现在有效的子概念。第三，如图句子4），它往往是DIF音响崇拜以检测其中的子概念列表开始或结束。因此，我们采用现阶段比较保守的方法，通过包括所有潜在的子概念成 $\bar{y}_{秒}$。

根据所使用的赫斯特模式，我们首先使用""作为分隔符提取candi-日期列表。为了 *持续* 元素，我们也使用

"与"和"或"打破它。由于诸如"与"和"或"可能会或可能不会是一个分隔符，我们把所有可能的候选人 $\bar{y}_{候}$。例如，例2中给出的句子3），我们有 $\bar{y}_3$ =

{IBM，诺基亚，宝洁，宝洁，宝洁}。

2.3.2超级概念检测

在案例 $|X_S|>1$，我们必须从不可能超概念 $X_{小号}$ 直到只有一个超级概念依然存在。我们使用超概念的检测概率方法。

让 $X_S=\{X_1, \cdots, X_{米}\}$。我们计算的可能性 $P(X_K|\bar{y}_S)$ 对于 $X_k \in X_{米}$。不失一般性，我们假设 X_1 和 X_2 有*最大*可能性和 $P(X_1|\bar{y}_S) \geq P(X_2|\bar{y}_S)$。我们计算似然比 $R(X_1, X_2)$ 如下然后我们挑选 X_1 如果该比值高于一个阈值：

$$R(X_1, X_2) = \frac{P(X_1|\bar{y}_S)}{P(X_2|\bar{y}_{多个}) = P(\forall s(X_{\in s}|R_s)|X_1)(X_2)}$$

由于候选子概念列表是众所周知的 *coor- dinate条款* 在文献中，这意味着它们在超级概念同样的IM portant，我们假设子概念 $\bar{y}_S$ =

$\{\bar{y}_1, \cdots, \bar{y}_N\}$ 独立给出的超级概念，并有

$$R(X_1, X_2) = \frac{P(X_1)\prod_{l=1}^{n} P(Y_{l米}|X_1)}{P(X_2)\prod_{l=1}^{n} P(Y_{l米}|X_2)}$$

我们计算上述比率如下：$P(X_{-米})$ 是有对的百分比 $X_{-米}$ 作为超级概念 Γ，和 $P(Y_{lj}|X_{-米})$ 是对的百分比 Γ 具有 $\bar{y}_J$ 给出子概念 $X_{-米}$ 是超级概念。当然，不是每一个 $(X_{-米}, \bar{y}_{j米})$ 出现在 Γ，在刚开始的时候尤其是 Γ 是小。这导致 $P(Y_{lj}|X_l) = 0$，

这使得我们无法计算出的比率。为了避免这种情况，我们让 $P(Y_{lj}|X_l) = \varepsilon$ 哪里 ε 是一个小的正数，当 $(X_{-米}, \bar{y}_{j米})$ 不 Γ。

作为一个例子，从实施例2中的句子1），我们得到 $X_{1米}$ = *{动物，狗}* 通过句法提取。直观上，likeli-罩 $P(动物|猫)$ 应远远高于 $P(狗|猫)$

在一个大的主体，因为它是像".....狗的句子不太可能如 *猫 ...* "存在，而喜欢的句子："..... 动物 如 *猫 ...* "相当普遍。其结果是，该比率 $R(动物，狗)$ 要大，和"动物"将被选作正确的超概念。

然而，上述方法不能ND科幻 *新* 超级概念。考虑了一句".....不是狗等家养动物 如 *猫 ...* "。如果我们还没有看到有"国内AN-imals"和"猫"许多句子构成，它是不可能的推断"国内AN-imals"作为一种可能的超概念。然而，常识告诉我们，*国内* 动物也是动物。以来 Γ 有一对（动物，猫），我们也可以得到一个大的可能性 $p(家养动物|猫)$ 并选择"六畜"的超级概念。在一般情况下，对于任何候选人的超级概念 X 在 $X_{小号}$ 不在 Γ，

我们带的作案科幻呃 X 并检查剩余的（更普遍）的概念 Γ 再次。这有助于我们收获从语料库更SPECI网络Ç理念和完善的召回。

2.3.3子概念检测

假设我们有identi网络编超级概念 $X_S=\{X\}$ 从一个句子。接下来的任务是连接第二从它的子概念 $\bar{y}_{的}$。在我们的工作中，子概念，检测是基于从句子中提取的特征。由于缺乏空间，在这里我们只专注于最重要的特点，从以下两个观察的。

Ø BSERVATION 1。*越接近备选子概念是对* 模式关键词，*越有可能是一个有效的子概念。*

事实上，一些提取方法（例如，[27]）仅取clos- EST一个以提高精度。例如，在应试PLE 2的句子3），*IBM* 最有可能是一个有效的子概念，因为它是右后 模式关键词 如，在句子4）*中国* 最有可能是一个有效的子概念，因为它是模式的关键字前右 和别的。

ØBSERVATION 2。*如果我们一定在一个备选子概念 K- 从个位置* 模式关键词 *是有效的，那么最有可能的候选从位置子概念 1 到位置 k - 1 也是有效的。*

在这里，一个备选子概念的位置编号相对于它的 *亲近* 到 *模式关键词。* 例如，在实施例2的句子3）中，第一个子概念是 *IBM*，第二子概念是 *诺基亚*，等等，而在句子4），在第一个子概念 *中国*，第二子概念是 *日本*，等等。

根据观察2，我们的策略是第一个网络第二最大范围，其中子概念都是有效的，然后解决范围内的上午biguity问题。

具体来说，我们科幻ND最大 k 使得似然 $P(Y_K|X)$ 高于阈值，其中 $\bar{y}_k$ 是备选子概念在 *K-* 从图案关键字个位置。但是，如果我们不能ND任何科幻 $\bar{y}_k$ 这SATIS网络上课的条件，那么我们假设 $K=1$，亲vided是 $\bar{y}_1$ 是 *胼，形成（* 例如，它不包含任何分隔符，如"与"或"或"），因为基于观察1，$\bar{y}_1$ 最有可能是一个有效的子概念。

例如，在句子4），我们可以正确决定的有效子概念的列表中结束 *澳大利亚*，如果的可能性 *Aus- tralia* 是不是以后考生的可能性更大 *中东，欧洲，* 和 *北美。*

然后，我们研究每个候选 $\bar{y}_1, \cdots, \bar{y}_k$。对于任何 $\bar{y}_{-世}$ 哪里 $1 \leq 一世 \leq K$，如果 $\bar{y}_{-世}$ 一点也不含糊，我们增加 $(X, Y_{-世})$ 至 Γ 如果它已不存在，否则我们增加计数 $N(X, Y_{-世})$。如果 $\bar{y}_J$ 是 *暧昧，* 也就是说，我们有位置的多种选择 $\hat{J}$，那么，我们需要决定哪一个是有效的。

假设我们有identi网络版 $\bar{y}_1, \cdots, \bar{y}_{J-1}$ 从位置1到位置 $\hat{J}$- 1 为有效子概念，并假设我们有两个candi-日期[3]在位置 $\hat{J}$，那是，$\bar{y}_J \in \{C_1, C_2\}$。我们计算likeli-罩比 *住宅（丙*$_1$，$C_2$）如下然后我们挑选 C_1 过度 C_2 如果该比值高于一个阈值：

$$住宅(丙_1, C_2) = \frac{P(C_1|X, Y_1, \cdots, \bar{y}_{J-1})}{P(C_2|X, Y_1, \cdots, \bar{y}_{J-1})}$$

和以前一样，我们假设 $\bar{y}_1, \cdots, \bar{y}_{J-1}$ 被赋予独立 X 和 $\bar{y}_J$，并且有：

$$住宅(丙_1, C_2) = \frac{P(C_1|X)\prod_{l=1}^{J-1} P(Y_{l米}|C_1, X)}{P(C_2|X)\prod_{l=1}^{J-1} P(Y_{l米}|C_2, X)}$$

这里，$P(C_1|X)$ 是对的百分比 Γ 哪里 C_1 是一个子概念，定 X 是超级概念，$P(Y_{l米}|C_1, X)$ 是likeli-罩那 $\bar{y}_{-世}$ 显示为有效的子概念，在一个句子里用 X 作为超级概念和 C_1 作为另一种有效的子概念。

作为一个例子，考虑句子3）例2中我们有三位候选多种选择，即 *宝洁Gam- BLE* 和 *普罗克特。* 直观地看，可能性为 *宝洁*

远远大于 *普罗克特* 因为机会 *普罗克特* 本身也是一种有效的公司名称是相当低的（即，它很可能非我们能遇到像".....公司的句子 如 *宝洁...* "）。因此，$p(宝洁|公司)$，

$p(IBM|宝洁公司，公司)$，和 $p(诺基亚|宝洁公司，公司)$ 应该都是比涉及的同行更大 *普罗克特。* 其结果是，该比率 $R(宝洁，宝洁)$ 要大，因此，我们选择"宝洁"为正确的子概念。

[3] 如前所述，如果我们有两个以上的候选人，我们选择这两个具有最大可能性。

3.分类学建筑

以前的步骤产生的一大组 *ISA* 对。每一对都表示在分类学的边缘。我们的目标是从这些个体的边缘构造一个分类。

3.1问题陈述

我们的分类为DAG（有向无环图）模型。在分类学的节点可以是一个 *概念* 节点（例如， *公司*），
或 *例* 节点（例如， *微软*）。一个概念包含了一组实例，可能一组子概念。边缘（ *U*，*V*）

连接两个节点 *ū* 和 *v* 意思是 *ū* 是的超级概念
诉 从实例节点区分概念节点是在我们的分类自然： *而不出边节点是实例，而其他节点的概念。*

创建一个图形出一套边的最明显的任务如下： *对于任何两个边各有一个节点具有相同的标签，我们应该把它们看成同一节点和两个边缘连接？* 考虑以下两种情况：

1.对于两个边缘 $\bar{E}_1$=（水果，苹果），$\bar{E}_2$=（公司，苹果），
应我们连接 $\bar{E}_1$ 和 $\bar{E}_2$ 节点"苹果"？

2.对于两个边缘 $\bar{E}_1$=（植物，树）$\bar{E}_2$=（植物，蒸汽涡轮机），
应我们连接 $\bar{E}_1$ 和 $\bar{E}_2$ 节点"植物"？

这个问题的答案都的问题显然是 *没有*，但我们如何决定对这些问题的？

显然，诸如"苹果"和"植物"可具有多个含义（感官）。因此，分类建设面临的挑战是这些感官来区分，并且具有相同的意义节点连接的边缘。我们进一步将问题分为两个子问题：1）集团通过概念幡然醒悟，即决定是否两个 *植物* 在第二个问题以上意味着同样的事情; 和

ii）在自己的感官，即组实例，决定是否两个 *苹果* 在第一个问题意味着同样的事情。我们认为，我们只需要解决的第一个子问题，因为一旦我们正确组由不同的感官所有的概念，我们可以通过它在概念层次中的位置，即确定一个实例的意义，它的意义取决于所有超级概念是这样。

我们攻击分两步分类建设问题。首先，我们确定的一些性质 *ISA* 对我们已经获得的。二，基于性能，我们引入两家运营商合并是属于同一感知节点，我们利用我们去定义网络运营商建立一个分类。

3.2感官

让 $X_{-世}$ 表示具有标签的节点 *X* 和感 *一世*。两个节点 $X_{-世}$ 和 X_J 是等价的当且仅当 *I = 学家* 对于边缘（*X*，*Y*），如果（$X_{-世}$，$\bar{y}_{J)的}$ 成立，则（$X_{-世}$，$\bar{y}_{J)的}$ 是可能的 *解释* 的（*X*，*Y*）。我们表示这是（*的x*，*y*）| =（$X_{-世}$，$\bar{y}_{J)的}$。给定一个边缘（*X*，*Y*），有3种可能的情况下，用于解释（*X*，*Y*）：

1.存在一个独特的 *一世* 和独特的 $\bar{J}$ 使得（*的x*，*y*）| =
（$X_{-世}$，$\bar{y}_{J)的}$。例如，（*行星*，*地球*）。这是最常见的情况。

2.存在一个独特的 *一世* 和多 $\bar{J}$ 的，使得（*的x*，*y*）| =
（$X_{-世}$，$\bar{y}_{J)的}$。例如，（*对象*，*植物*）。

3.存在着多个 *一世* 的和多 $\bar{J}$ 的，使得（*的x*，*y*）| =
（$X_{-世}$，$\bar{y}_{J)的}$。这种情况在实践中是非常罕见的。最后，这是不可能存在多个 *一世* 不过是一个独特的
$\bar{J}$ 使得（*的x*，*y*）| =（$X_{-世}$，$\bar{y}_{J)的}$。

3.3属性

我们透露了一些重要的特性 *ISA* 对我们通过赫斯特模式而获得。在我们的讨论中，我们用下面的句子作为我们当前实例。

ËXAMPLE 3。 *正在运行的例子。*

一个）...植物 如 树木 和 草...

B）...植物 如 树木，草 和 草药 ...

C）...植物 如 蒸汽涡轮机，泵，和 钢炉...

d）...生物 如 植物，树木，草 和 动物 ...

E）...事 如 植物，树，草，泵，和 钢炉...

P ROPERTY 1。 *让 S =｛（X，Y₁），⋯，（X，Yₙ）｝是* ISA *从句子派生对。然后，所有的 X 的中 小号 具有独特的意义，那就是，存在一个独特的 一世 使得（X，Y_{J)的}）=（$X_{-世}$，$\bar{y}_{J)的}$）适用于所有 1 ≤ J ≤ ñ。*

直观地说，这意味着这样的句子"...植物 如 树木 和 钢炉..."是极为罕见的。换句话说，在所有的超级概念 *ISA* 从单句对具有同样的意义。例如，在句子a）中，字的感官 厂S IN（*植物，树木*）和（*植物，* 草）是相同的。在此之后丙erty，我们表示 *ISA* 从一个句子对为｛（$X_{-世}$，$\bar{y}_1$），⋯，（$X_{-世}$，$\bar{y}_N$）｝

强调所有 *X* 的具有相同的意义。

P ROPERTY 2。 *设｛（$X_{-世}$，$\bar{y}_1$），⋯，（$X_{-世}$，$\bar{y}_{米}$）｝从一个句子表示对，和｛（X_J，$\bar{z}_1$），⋯，（X_J，$\bar{z}_N$）｝从另一个句子。如果｛$\bar{y}_1$，…$\bar{y}_{米}$｝和｛$\bar{z}_1$，…$\bar{z}_{N}$｝是相似的，那么它很有可能 $X_{-世}$ 和 X_J 是等价的，也就是说，I = 学家*

考虑句子a）和b）实施例3中的一组子概念有一个大的重叠，因此，我们的结论是该单词的感官 *植物* 在两个句子是相同的。同样的事情，不能说对句子b）和c），为无相同的子概念中找到。

P ROPERTY 3。 *设｛（$X_{-世}$，Y），（$X_{-世}$，$\bar{u}_1$），⋯，（$X_{-世}$，$\bar{u}_{米}$）｝表示从一个句子获得对，和｛（$\bar{y}_K$，v_1），⋯，（$\bar{y}_K$，v_N）｝从AN-其他句子。如果｛$\bar{u}_1$，$\bar{u}_2$，…$\bar{u}_{米}$｝和｛v_1，v_2，…v_{N}｝是相似的，那么它很有可能是（$X_{-世}$，Y）| =（$X_{-世}$，$\bar{y}_K$）。*

这意味着词 *植物* 句子d）具有相同的意义，因为这个词 *植物* 在句子中的），因为他们的子概念有相当多的重叠。这不是句子d）和c）真。出于同样的原因，这个词 *植物* 在句子E）可以被解释为感 *植物* 在a）和c）在同一时间。

3.4节点合并操作

根据这三个特性，我们马上就可以开发一些机制来参加由发自内心的终端节点的边。

首先，根据性质1，我们知道，每一个超概念的 *ISA* 从单句得出对具有同样的意义。因此，我们参加这样 *ISA* 超概念节点上对（见无花果URE 1）。我们把从单一的句子中所获得的分类

本地分类。 本地分类与根 $X_{-世}$ 被表示为 $\bar{T}_{-世}$ $_{X_0}$
其次，根据性质2，给定的扎根两个本地分类 $X_{-世}$ 和 X_J，如果两个分类的子节点演示相当的相似性，我们执行 *水平合并*（见无花果URE 2）。

第三，基于物业3，鉴于在扎根两个本地分类
$X_{-世}$ 和 $\bar{y}_K$，如果 $X_{-世}$ 有一个子节点 *Y*，和两个分类的子节点演示相当的相似性，我们将二者合并

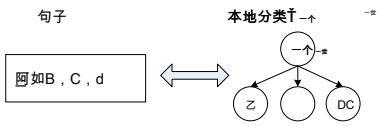


图1：从单句到本地分类

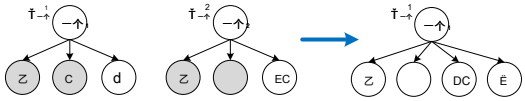
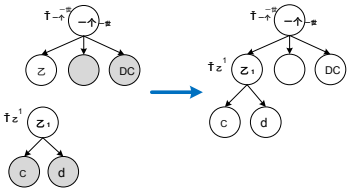


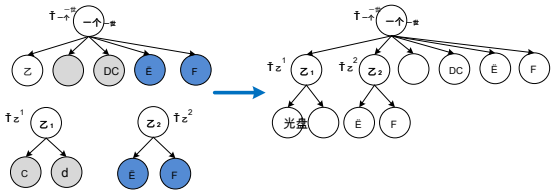
图2：横向合并

在节点上的分类法 *年*。 我们称之为 *垂直合并*， 并且在图3（a）中示出。

用于垂直合并的特殊情况如图3（b）所示。都 $\hat{T}_1$ z和 $\hat{T}_2$ z与垂直可合并 $\hat{T}_{-世}$ $-个$，作为孩子节点 $\hat{T}_1$ z和 $\hat{T}_2$ z有与孩子相当大的重叠节点 $\hat{T}_{-世}$ $-个$。然而，子节点 $\hat{T}_1$ z和 $\hat{T}_2$ z没有相当大的重叠，因此它仍然是可能的，它们代表两种意义。其结果是两个子分类，如图3（b）所示。



（A）单一取向感



（b）中的多比对感

图3：垂直合并

3.5相似度函数

假设我们使用 *儿童（T）* 表示本地分类的子节点 *吨*。 鉴于两个本地分类 $\hat{T}_1$和 $\hat{T}_2$，躺在水平和垂直合并操作的中央步骤是检查的重叠 *儿童（T_1）* 和 *儿童（T_2）*。在一般情况下，我们可以去网络连接NE的相似功能 $F(A,B)$ 这样两套 *一个* 和 *乙*

是 *类似（* 表示为 $辛(A,B)$ ）如果 $F(A,B) \geq \tau(A,B)$， 哪里 $\tau(A,B)$ 是一些prespeci音响编阈值函数。

然而，选择 $F(A,B)$ 是不是任意的。相似去网络连接在一个相对的方式定义，诸如捷卡度量，常可导致不合理的结果。请看下面的例子：

让 $A=\{$微软，IBM，HP$\}$， $B=\{$微软，IBM，英特尔$\}$，$C=\{$微软，IBM，HP，EMC，英特尔，谷歌，苹果$\}$，并燮提出的是的超级概念 *一个*，*B*， 和 *C* 是"IT公司"。让 $\hat{J}(X,Y)=|x_n y|/|x_u y|$ 是的Jaccard相似 *X* 和 $\hat{y}$。 然后 $\hat{J}(A,B)=2$ $\bar{4}=0.5$，而 $\hat{J}(A,C)=3$ $\bar{7}=0.43$。如果我们设置相似性阈值是0.5，则 *一个* 和 *乙* 被认为是SIM- ILAR，而 *一个* 和 *C* 不是。其结果是，水平方向的合并可以应用于 *一个* 和 *B*，但不 *一个* 和 *C*。 这是因为荒谬 *一个* 是 *子集* 的 *C*。

因此，我们重点关注 $F(A,B)$ 通过将测量 *绝对* 重叠 *一个* 和 *B*. 此外，对于房产2和3中，多个重叠的证据，我们有，更多的CON组fi凹痕我们是在形成的合并操作per-。因此，我们要求 $F(A,B)$ 符合下列财产：

P ROPERTY 4。 *如果 A，A'，B，和 乙'任何组ST 一个⊆一个'和 乙⊆乙'，然后 SIM（A，B）⇒辛（A'，乙'）。*

那么最简单的方法可能是去连接的NE $F(A,B)=|A\cap C|$ 然后让 $\tau(A,B)$ 等于某个常数 δ。 在此设置下，我们有 $FA',乙') \geq F(A,B)$。于是 $F(A,B) \geq \delta \Rightarrow FA',乙') \geq \delta$，亦即 $SIM(A,B) \Rightarrow 辛(A',乙')$。

3.6算法

算法2总结分类重刑建筑的框架。整个过程可以分为三个阶段，即 *本地分类建设* 阶段（线3-5），该 *水平分组* 阶段（线6-10），和 *垂直分组* 阶段（线11-19）。我们首先创建从每个句子本地分类 *小号* $\in S$.

然后，我们执行的本地分类，其根节点具有相同的标签水平合并。在这个阶段，小地方的分类法将合并，形成较大的。最后，我们进行本地分类法，其根结点有不同的标签，垂直合并。

算法2： 分类建设

输入： S： 该组句子每一个包含若干 *ISA* 对。

输出： T： 该分类图。

```
1 让 T̂ 是集本地分类的;
2 牛逼←∅;
3的foreach S={(X_世, ŷ_1), ..., (X_世, ŷ_N)} ∈ 小号 做4
      添加本地分类 T̂_世 x成 吨;
5端6的foreach T̂_世 x∈T， T̂_J x∈T̂做
7     如果 SIM(儿童(T_世 x)，儿童(T_J x)) 然后
8        HorizontalMerge(T_世 x，T̂_J);
9     端10的端部
11的foreach T̂_世 x∈T̂做
12    的foreach ŷ∈儿童(T_世 x) 做
13       的foreach T̂_米 ŷ∈T̂做
14          如果 SIM(儿童(T_世 x)，儿童(T_米 Y)) 然后
15             VerticalMerge(T_世 x，T̂_米);
16          结束
17       结束18
      端19端部20 让这样连接的图形是 吨;
21回 吨;
```

在算法2，我们的垂直分组之前执行水平分组。这个命令是没有必要的。作为定理1个使然，合并操作的任何序列，最终导致同样taxon- OMY。然而，如定理2表明，当水平分组垂直分组，这是更加理想的做之前被最小化的合并操作的总数量。两个定理的证明可以在[40]中找到。

THEOREM 1。 *让 T̂ 是一组本地分类法。让 Ø$_\alpha$ 和 Ø$_\beta$ 是水平和垂直合并歌剧蒸发散上的任何两个序列 吨。 假设没有进一步的操作可以进行上 T̂ 后 Ø$_\alpha$ 要么 Ø$_\beta$。然后，在执行后的音响最终图形 Ø$_\alpha$ 并执行之后的音响最终图形 Ø$_\beta$ 是相同的。*

Ť HEOREM 2。 ***让 Ø 是集合OP-操作的所有可能序列的，并让 M = 分{| Ø| ：o ∈Ø}。假设 Ø ₒ是执行所有可能的水平的序列合并第一个和所有可能的垂直合并下，则|Ø ₒ| = M.***

3.7相关工作

自动分类学结构已在文献[6，28，8，21，32，35，26]被广泛地研究。早期的工作[6，28，8]佛cused对诱导闭域的分类法，或曾经存在延伸荷兰国际集团开放域的分类法如的Cyc [21]和WordNet的[32]。在小分类前者结果CON组fi奈德在一个特定的ÇDO-主，**而后者通过添加更多的丰富了分类学 *命名实体*。换句话说，只有数 *实例* 的增加，而不是数量 *概念* 和 *ISA* 概念之间的关系。更多的讨论可以在[26]中找到。**

从头开始创建开放域分类一般是 - lieved比上述任务更具挑战性。对这项工作最无糖**表工作isWikiTaxonomy [26]和YAGO [35]。他们都试图获得一个分类 [4] 从维基百科**

类别。 **虽然这些工作报告的精度高（不外加攝由于来自维基百科更清洁的数据），他们有两个基本的限制。首先，维基百科类别 *the- MATIC主题* 用于classifyWikipedia的文章，这是从的术语相当昼夜温差ferent *概念* 通过分类学德音响nition提及。例如，WikiTaxonomy包含"概念"，如" *雅典的历史* " 要么 " *加拿大地质* "这是真正的主题，而不是概念或范**畴。二，分类的覆盖完全依靠维基百科，这仍然是有限的与游离碱或Probase相比的覆盖范围。

据我们所知，目前是在非盟tomatically诱导分类没有现成的工作 *网络规模* 类似于本文的努力。突出web规模信息提取系统，如KnowItAll [12]和TextRunner [2] *提取* 该 *ISA* 从网页上双，但赶不上上的概念感官构建从这些对有区别的分类法的。

4.概率分类学

知识并不是非黑即白。删除所有的不确定性不仅是不可能的，而且是有害的。我们的理念是生活在嘈杂的数据，使物尽其用。在本节中，我们亲构成概率框架模型在Proba**se知识。这个框架实际上由两个部分组成：一个企业的联合概率 *ISA* 一对，也被称为 *合理性;* 和一个概念，其实例之间的条件概率，又称为 *典型性。***

4.1合理性

对于每一个要求 *Ė* 在Probase我们使用 *P（E）* 来表示prob-能力，这是真的。在许多情况下，有真假之间没有明显的区别; 并且在几乎所有的情况下，来自外部数据源的信息**是不可靠的，不一致的，或错误的。因此，我们解读 *P（E）* 作为 *合理性* 的 *E.* 在本文中，我们重点关注 *ISA* 的关系，也就是说，每个权利要求 *Ė* 是的一个权利要求 *ISA***

之间的关系 *X* 和 *Y*，和 *P（E）= P（X，Y）*。

推导 *P（E）*，我们认为外部信息作为证据

***E.* 具体来说，假设 *Ė* 从一组的句子或EV-idence {衍生 *小号₁*，…，*小号ₙ*} 在网上。假设每**件证据

***小号₋世* 与相关联的概率 *p₋世* 那些重新FL学分的信仰或证据CON组连接的信心。在这里，我们采用简单的喧闹，或模型。索赔 *Ė* 是假的当且仅当在每一件证据 *小号₁*，…，*小号ₙ* 是假的。由于不同的句子是从不同的网页提取**

这是分别由不同的人创造的，我们假设的证据是独立的，并有

$$P(X,Y) = 1 - p(\bar{E}) = 1 - p(\bigwedge_{i=1}^{n} \text{小号}_i) = 1 - \prod_{i=1}^{n} (1 - p_{\text{小号}_i}). \quad (1)$$

更复杂的模型（如瓮模型[11]），可用于似然性。由于缺乏空间，我们重点对那些喧**闹，或模型的讨论。该模型具有良好的可扩展性。这是很容易集成新的证据，包括 *负* 证据。负证据声称，*A不是B*，这有效地降低了对如权利要求的合理性在于 *A是B.* 例如负证**据包括所述 *部分* 关系，例如" *乙由...构成* 一个，*C*，**

和...* "纳入负面证据 *小号₋世* 概率 *p₋世

进入合理性是直接的。我们只需更换因素

1 - *p₋世* 同 *p₋世* 公式。（1）。

剩余的问题是如何获得 *p₋世* 证据 *小号₋世*。我们CON组代尔两个因素。第一，*p₋世* 可能取决于Informa公司和灰源的类型（例如，我们认为证据来自纽约时报到来比那些从公共论坛更可信）。第二，

***p₋世* 可能取决于信息提取过程是如何CON组fi凹痕是当它identi音响ES证据 *小号₋世* 在文本中。在我们的工作中，我们煤焦acterize各 *小号₋世* 由一组功能 *F₋世*，如：i）所述PageRank得分的页面的从中 *小号₋世* 是萃取 ii）用在赫斯特图案 *小号₋世*; III）的句子用数字 *X* 作为超概念;**

iv）句子的与数 *ÿ* 作为子概念; v）的的子概念数 *小号₋世*; vi）的位置 *ÿ* 在 *小号₋世*; 等等。然后，原样suming独立的功能，我们可以应用朴素贝叶斯推导

***p₋世*。具体来说，我们有**

$$p_1 = P(S_{\text{到}} F_i) = \frac{P(S_{\text{小号}_i}) \cdot \prod_{F \in F_{\text{小号}_i}} P(F|S_{\text{小号}_i})}{\sum_{\text{小号}_i \in \{\text{小号}_1, \ldots, \text{小号}_n\}} P(S) \cdot \prod_{F \in F_{\text{小号}_i}} P(F|S)}。 \quad (2)$$

要了解该模型中，我们使用共发现建立一个训练集。给定一对（*X*，*Y*），如果两个 *X* 和 *ÿ* 出现在WordNet中，并有从路径 *X* 至 *ÿ* inWordNet，那么我们考虑（*X*，*Y*）作为一个 *正*

例; 如果两个 *X* 和 *ÿ* 出现在WordNet的，但是从没有路径 *X* 至 *Y*，然后我们考虑（*X*，*Y*）作为一个 *负* 例。

4.2典型性

直观地说，知更鸟是更 *典型* 概念 *鸟* 比OS-trich，而 *微软* 更 *典型* 概念 *公司* 比 *XYZ 公司*.. 然而，在所有现有的分类，概念里面的实例被看作 *相等* 典型。例如，*公司*

概念的游离碱含有约79000情况下没有办法来衡量他们的典型性。关于典型性的信息是许多应用，如在第1节介绍是必不可少的。

在本文中，我们提出了一个典型性的措施。典型性存在于两个方向：一个实例的概率**给出一个概念 *T（我|X）***

（又名实例化）; 和概念的概率给出一个实例

***T（X|i）*（又名抽象）。我们专注于前者这里，后者可以通过简单的方式贝叶斯法则来**计算。

***T（我|X）* 取决于两个因素：1）的数量 *N（x，I）* 的支持，如权利要求证据（*X*，*I*）; 和ii）的似然性 *P（X，I）* 权利要求的（*的x，i）中*，作为德科幻定义公式。（1）。然后**我们去连接NE的典型性

一世至 *X* 如：

$$T(\text{我}|X) = \frac{N(x,I) \cdot P(X,I)}{\sum_{\text{一世}_x \in \text{一世}_X} N(X,I_j) \cdot P(X,I_j)}. \quad (3)$$

哪里 *一世ₓ* 是集合概念的实例 *X*。

一个问题是，除了直接对（*的x，i）中*，*我* 也可以是一个概念的一个实例 *ÿ* 这是的子概念 *X*。例如导出，除了声称 *微软* 是概念的一个实例

***公司*，也有人声称在Probase那 *微软* 是概念的IN-姿态 *IT公司* 和 *大公司*。这些说法**

这是分别由不同的人创造的，我们假设的证据是独立的，并有

$$P(X,Y) = 1 - p(\bar{E}) = 1 - p(\bigwedge_{i=1}^{n} \text{小号}_i) = 1 - \prod_{i=1}^{n} (1 - p_{\text{小号}_i}). \quad (1)$$

更复杂的模型（如瓮模型[11]），可用于似然性。由于缺乏空间，我们重点对那些喧**闹，或模型的讨论。该模型具有良好的可扩展性。这是很容易集成新的证据，包括 *负* 证据。负证据声称，*A不是B*，这有效地降低了对如权利要求的合理性在于 *A是B.* 例如负证**据包括所述 *部分* 关系，例如" *乙由...构成* 一个，*C*，**

和...* "纳入负面证据 *小号₋世* 概率 *p₋世

进入合理性是直接的。我们只需更换因素

1 - *p₋世* 同 *p₋世* 公式。（1）。

剩余的问题是如何获得 *p₋世* 证据 *小号₋世*。我们CON组代尔两个因素。第一，*p₋世* 可能取决于Informa公司和灰源的类型（例如，我们认为证据来自纽约时报到来比那些从公共论坛更可信）。第二，

***p₋世* 可能取决于信息提取过程是如何CON组fi凹痕是当它identi音响ES证据 *小号₋世* 在文本中。在我们的工作中，我们煤焦acterize各 *小号₋世* 由一组功能 *F₋世*，如：i）所述PageRank得分的页面的从中 *小号₋世* 是萃取 ii）用在赫斯特图案 *小号₋世*; III）的句子用数字 *X* 作为超概念;**

iv）句子的与数 *ÿ* 作为子概念; v）的的子概念数 *小号₋世*; vi）的位置 *ÿ* 在 *小号₋世*; 等等。然后，原样suming独立的功能，我们可以应用朴素贝叶斯推导

***p₋世*。具体来说，我们有**

$$p_1 = P(S_{\text{到}} F_i) = \frac{P(S_{\text{小号}_i}) \cdot \prod_{F \in F_{\text{小号}_i}} P(F|S_{\text{小号}_i})}{\sum_{\text{小号}_i \in \{\text{小号}_1, \ldots, \text{小号}_n\}} P(S) \cdot \prod_{F \in F_{\text{小号}_i}} P(F|S)}。 \quad (2)$$

要了解该模型中，我们使用共发现建立一个训练集。给定一对（*X*，*Y*），如果两个 *X* 和 *ÿ* 出现在WordNet中，并有从路径 *X* 至 *ÿ* inWordNet，那么我们考虑（*X*，*Y*）作为一个 *正*

例; 如果两个 *X* 和 *ÿ* 出现在WordNet的，但是从没有路径 *X* 至 *Y*，然后我们考虑（*X*，*Y*）作为一个 *负* 例。

4.2典型性

直观地说，知更鸟是更 *典型* 概念 *鸟* 比OS-trich，而 *微软* 更 *典型* 概念 *公司* 比 *XYZ 公司*.. 然而，在所有现有的分类，概念里面的实例被看作 *相等* 典型。例如，*公司*

概念的游离碱含有约79000情况下没有办法来衡量他们的典型性。关于典型性的信息是许多应用，如在第1节介绍是必不可少的。

在本文中，我们提出了一个典型性的措施。典型性存在于两个方向：一个实例的概率**给出一个概念 *T（我|X）***

（又名实例化）; 和概念的概率给出一个实例

***T（X|i）*（又名抽象）。我们专注于前者这里，后者可以通过简单的方式贝叶斯法则来**计算。

***T（我|X）* 取决于两个因素：1）的数量 *N（x，I）* 的支持，如权利要求证据（*X*，*I*）; 和ii）的似然性 *P（X，I）* 权利要求的（*的x，i）中*，作为德科幻定义公式。（1）。然后**我们去连接NE的典型性

一世至 *X* 如：

$$T(\text{我}|X) = \frac{N(x,I) \cdot P(X,I)}{\sum_{\text{一世}_x \in \text{一世}_X} N(X,I_j) \cdot P(X,I_j)}. \quad (3)$$

哪里 *一世ₓ* 是集合概念的实例 *X*。

一个问题是，除了直接对（*的x，i）中*，*我* 也可以是一个概念的一个实例 *ÿ* 这是的子概念 *X*。例如导出，除了声称 *微软* 是概念的一个实例

***公司*，也有人声称在Probase那 *微软* 是概念的IN-姿态 *IT公司* 和 *大公司*。这些说法**

---

[4] YAGO实际上是与包括分类Infor公司，mation更一般的本体。

也应被视为额外的证据， *微软* 是典型的 *公司*。

将这些 *间接* 证据还有，我们重新连接的NE式。（3）为：

$$T(\text{我}|X)=\frac{\sum_{\bar{y}\in d(x)\text{的}}\frac{}{P(X,Y)}\cdot N(Y,I)\cdot P(\text{义})}{\sum_{\text{一度-e-度x}}\sum_{\bar{y}\in d(x)\text{的}}\frac{}{P(X,Y)}\cdot N(Y,I_{\cdot})\cdot P(\text{义}_{\cdot})}, \quad (4)$$

哪里 $\frac{}{P(X,Y)}$ 是的合理性是 $\bar{y}$ 是一个子概念或 *descen-唐太斯* 的概念 X，即，存在来自至少一个路径 X 至 年。我们用 $d(x)$ 的 来表示所有的子概念和后代的概念 X。特别是，$X\in d(x)$ 的 我们去科幻NE $\frac{}{}$

$$P(X,X)=1。$$

剩下的问题是计算 $\frac{}{}$ $P(X,Y)$。形式上，让 P_{xy} 是有至少一个路径从事件 X 至 年。假设 *家长（Y）* 是一组的直接超概念 年。对于每一个 $\bar{z}\in$ *家长（Y）* 让 $P_{\bar{z}}$ 是事件 $\bar{y}$ 是的直接子概念 $\bar{z}$ 并且存在从至少一个路径 X 至 Z，即 $P_{\bar{z}}=P_{zy}\wedge P_{xz}$。假设独立性 P_{zy} 和 P_{xz}，我们有

$$p(P_{\bar{z}})=p(P_{zy}\wedge P_{xz})=p(P_{zy})\cdot p(P_{xz})=P(Z,Y)\cdot p(P_{xz})。(5)$$

此外，假定的独立性 $P_{\bar{z}}$ 对于 $\bar{z}\in$ *家长（Y）*

我们再有

$$p(P_{xy})=p(\vee_{\bar{z}\in\text{家长}(Y)}P_{\bar{z}})=1-p(\wedge_{\bar{z}\in\text{家长}(Y)}\bar{P}_{\bar{z}})$$

$$=1-\prod_{\bar{z}\in\text{家长}(Y)}(1-p(P_{\bar{z}}))。 \quad (6)$$

代式。（5）代入方程。（6），我们得到

$$P(\text{的}x,y)=P(P_{xy})=1-\prod_{\bar{z}\in\text{家长}(Y)}(1-P(Z,Y)\cdot\frac{}{P(X,Z)}) \quad (7)$$

算法3： 计算 $\frac{}{P(X,Y)}$

输入：T：该分类

输出：$\Gamma=\{\frac{}{P(\text{的}x,y)}\}$：其中存在从路径 X 至 $\bar{y}$

```
1 Γ ← ∅;
2 k ← 1;
3 大号1 ← 概念 T̄ 没有父母的;
4 ，而 大号k ≠ ∅ 做5
      的foreach ȳ ∈ 大号k 做6
            如果 父（Y）= ∅ 然后7
                  加 P(Y,Y) = 1 至 Γ;
8           其他9
                  的foreach X ∈ 祖先（Y）做
10                      P(X,Y) =
                        1 − ∏_{z̄∈家长(Y)} (1 − P(Z,Y) · P(X,Z));
11                      加 P(X,Y) 至 Γ;
12                结束
13          结束14
      结束15
      k ← K + 1;
16    大号k ← 概念 T̄ 不在 U_{l=1 大号 −度} 但所有
      父母 U_{l=1 大号 −度}
17月底18回 Γ;
```

Algorithm3描述了动态规划方法的COM普泰{ $\frac{}{}$ $P(\text{的}x,y)\}$。它遍历一个自上而下的方式的分类。每次计算一段时间 $\frac{}{P(X,Y)}$ 在第10行，所需要的

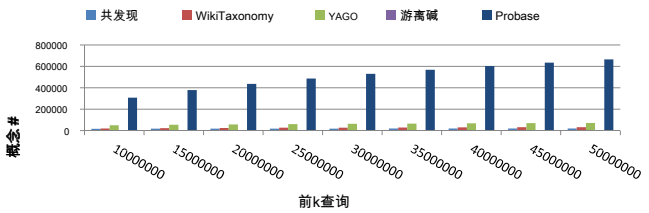


图4：在分类法相关概念号码

$P(X,Z)$ 的是保证已经计算。从派生典型性 $\frac{}{}$ $P(X,Y)$ 是简单的，因此这里未示出。

4.3相关工作

概率的方法已被利用在一些先前的工作[28，32，12，2，36]。在[28]和[32]的基础上，STA-tistical学习框架在分类学感应被使用。在KnowItAll [12]和TextRunner [2]，CLASSI音响ERS用于分配置信得分的 *ISA* 对提取。我们的工作从这些不同的两种原样pects。首先，在以前的工作中，概率仅用于 *中* 提取或分类学诱导阶段，而音响最终输出分类学仍然是 *确定性*。例如，在KnowItAll和TextRunner使用的置信度得分仅用于网络连接1-的TeringBay不正确的目的 *ISA* 对。其次，虽然在KnowItAll和TextRunner的置信度得分可以共享一些相似性与合理性，我们去这里科幻定义，他们的得分的语义是不明确的。此外，我们的重点放在塑造的典型性新颖而且从来没有明确的处理。尽管[36]杠杆scor- ING式（仍 *中* 诱导期），在语义我们典型性德网络nition有点重叠，其公式不会AP- plicable一般的分类，因为它们严重依赖于Infor公司，mation SPECI网络C到维基百科。他们还没有考虑不确定性（如，合理性）因素，我们认为这是必要的任何自动分类推理的框架。在实际文本理解任务的典型性的effective-内斯已经被我们最近的工作[34，39，37]证明。

5.实验评价

建议的分类推理的框架是使用服务器集群上实现 *的map-reduce* 模型。我们用7小时10个机ND所有科幻 *ISA* 对，然后4小时30台机器构建分类。我们还呼吁三位一体[29，30]图形数据库系统主机Probase。由于篇幅CON组straints，只提供结果的亮点。读者重新ferred到 http://research.microsoft.com/probase/

完整的实验结果。

我们提取326110911个句子含有语料库1679189480个网页。据我们所知，我们的语料规模是数量级比以前已知的最大的语料库[27]大一个数量级。推断的分类有2653872个不同的概念，不同的16218369概念实例对，

4539176不同概念子概念对（20757545双总共）。接下来，我们分析了概念空间和特征 *ISA* Probase的关系空间，并简要地评价ÿ几个应用程序，充分利用典型性的概念。

5.1概念空间

鉴于Probase有更多的概念，比任何其他分类，合理的问题是他们是否理解文本更有效。我们衡量effective-的一个方面，通过检查网络搜索查询Probase的概念覆盖内斯在这里。我们去科幻定义一个概念是 *相关的*，如果在网络查询出现至少一次。我们分析了Bing的查询日志从两年的期间，通过频率递减顺序排序的查询，

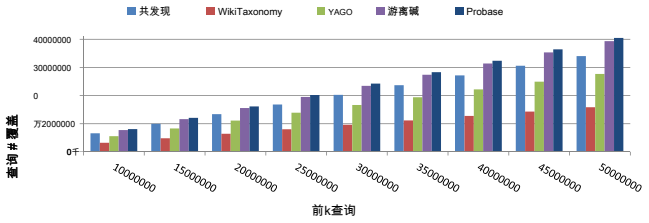


图5：前50万次查询的分类覆盖面

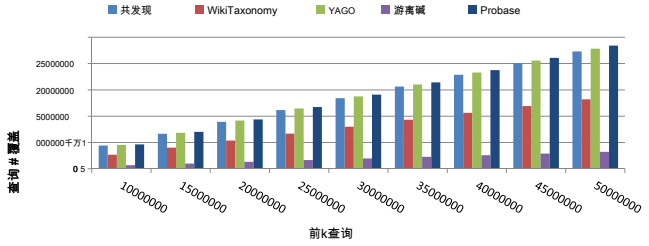


图6：前50万次查询的概念覆盖

和计算的在Probase相关概念和其它四个通用分类的数量，即 *共发现*，*WikiTaxon- OMY [26]* *YAGO*，和 *游离碱*，相对于前50万次查询。图4显示了结果。

总共664775个概念被认为是Probase有关，相比之下，只有70,656在YAGO。用户网**络查询具有已知的良好 *长尾巴* 分配。尽管少数基本CON组cepts的（例如，*公司*，*城市*，*国家*）**代表常识性知识非常频繁地出现在用户的查询，互联网用户不要说其他少为人知的概念。Probase做一个更好的工作在长尾捕捉这些概念，因此具有理解这些用户查询一个更好的机会。

接下来，我们测量 *分类覆盖* 由Probase查询。查询被认为是 *覆盖* 通过分类，如果查询包含至少一个 *概念或实例* 该分类内。[五] 图5的COM剥通过Probase分类查询，对其他四个分类的覆盖范围。Probase优于上的前10万元覆盖到前50万次查询其他人。总之，Probase覆盖

40517506（或81.04%），前50万次查询的。我们进一步测量 *概念覆盖*，其含有至少一个查询的数量 *概念* 在分类。图6中由Probase针对其他四个分类法进行比较的概念覆盖。同样，Probase优于所有其他人。需要注意的是，尽管游离碱呈现Probase媲美分类覆盖率在图5中，其概念范围要小得多。

总之，具有较大的概念空间，Probase表现出捕获在用户查询中隐含的语义能力较强。那么预计Probase可以解释这些查询的有用工具。最近的研究[39]通过查询解释利用Probase进一步证明了有效性。

5.2 ISA关系空间

有两种 *ISA* 在Probase关系：所述concept-子概念关系，其在层次结构中连接内部节点的边，和概念实例关系它们连接的叶节点的边缘。

表4 Probase的概念子概念关系空间与其它分类法进行比较。该 *水平* 一个概念是定义为从它的最长路径的到叶节点的长度（即，一个实例）去网络连接。表4表明，即使与MAG-的顺序

[五] **我们说"盖"的意思是，分类 *有助于* 于该查询的非derstanding因为它理解在查询的至少**一个词。

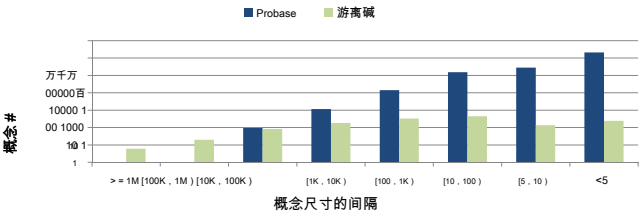


图7：在Probase和游离碱概念尺寸分布

nitude较大的概念数，Probase仍具有相当的复杂层次的其他分类。唯一的例外是游离碱**，其在这些测定量度表现微不足道的值，因为它没有 *ISA* 它在所有概念之间的关系。**

	排名 *ISA* 对	儿童平均#的	平均# 父母	平均水平	最高水平
共发现	283070	11.0	2.4 1.265		14
WikiTaxonomy	90739	3.7	1.4 1.483		15
YAGO	366450	23.8	1.04 1.063		18
游离碱	0	0	0	1	1
Probase	4539176	7.53	2.33 1.086		7

表4：概念-子概念关系空间

我们也比较Probase和游离碱的概念实例关系。我们选择的游离碱，因为它是与实例**空间（24483434概念实例对）规模相媲美的唯一的现有分类。我们去科幻NE *概念大小*为实例的数量直接下一个概念节点。**

图7比较了在Probase和游离碱概念尺寸的分布。尽管游离碱的重点，如"一些非常流**行的概念 *服踪*"和" *书*"有超过两百万的情况下，Probase具有小尺寸的概念，更多的媒体**。事实上，在游离碱排名前10位的概念包含17174891概念实例对，或者它拥有的所有成对的70%。相反，在Probase顶部10的概念只包含727136双，或者其总的4.5%。因此，Probase提供了各种主题的更广阔的覆盖范围，而游离碱是SPECI科幻ç主题更多的信**息。在另一方面，在游离碱的大概念，就像实例 *书* 大多来自SPECI网络Ç网站如亚马逊**，可以很容易地合并到Probase。

为了估计所提取的正确性 *ISA* 在Probase对，我们创建的各领域40个概念的基准数据集。由于篇幅所限，我们向他们展示的20在表5中的COM plete基准可以在本文中[40]的**更长的版本中找到。这个概念从大小21个实例变化（对 *飞机模型*）**

到85391（用于 *公司*），与917的基准用的也有报道在信息提取研究[25]的概念和领域覆盖同样数量的中位数。对于每个概念，我们随机选择了50个实例/子概念，并要求法官的人，以评估其正确性。

图8示出的结果。在基准所有对的平均精度为92.8%，其性能优于精度从像知道 - ItAll [12]（64平均%），NELL [7]（74%）和其他TextRunner以前显着信息提取框架报道[2]（80%平均）。这是不公平的直接比较我们基于维基百科的框架结果一样WikiTaxonomy [26]（86%）和YAGO [35]（95%），其数据来源是干净多了。Never- theless，只有YAGO比Probase一个更好的整体精度。

我们还研究了提取框架的每个迭代。图9示出的累计数量的 *ISA* 对和概念每轮迭代后萃取。两条曲线迅速成长起来的第一个几个回合，然后饱和的引导过程收敛。一个有趣的现象是，最大增益实际上发生在第二轮，而不是第一个之一。这是因为在第一个回合，在许多句子歧义

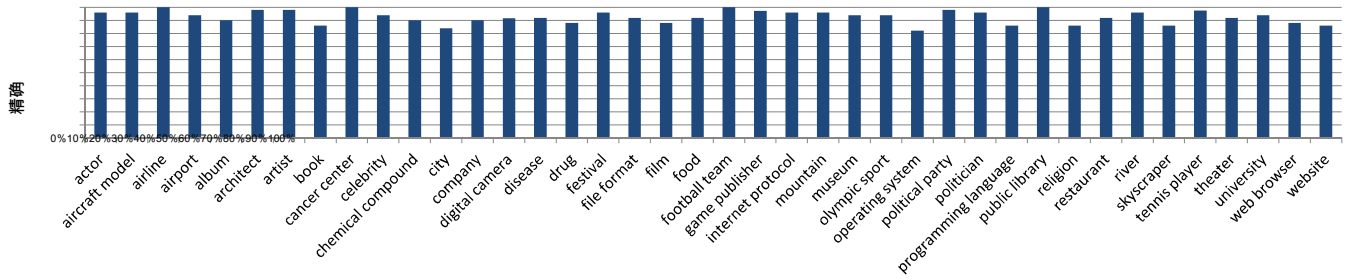


图8：提取对精密

概念（实例＃）典型实例	
演员（3466）	汤姆·汉克斯，马龙·白兰度，乔治·克鲁尼
飞机模型（21）	空中客车A320-200，苹PA-32，山毛榉-18
航空公司（1221）	英国航空公司，DeltaE值
专辑（1938）	惊悚，大平静的，肮脏的心灵
癌症中心（55）	福克斯蔡斯，护理联盟，达纳 - 法伯
化学化合物（308）	二氧化碳，菲，一氧化碳
公司（85391）	IBM，微软，谷歌
疾病（8784）	艾滋病，阿尔茨海默病，衣原体
节日（3039）	圣丹斯电影节，圣诞节，排灯节
音响LM（13402）	银翼杀手，星球大战，独领风骚
橄榄球队（59）	皇家马德里，AC米兰，曼联
因特网协议（168）	HTTP，FTP，SMTP
博物馆（2441）	卢浮宫，史密森，古根海姆博物馆
操作系统（86）	Linux，Solaris和的Microsoft Windows
政治家（953）	奥巴马，布什，布莱尔
公共图书馆（39）	哈林盖，加尔各答，诺维奇
餐厅（4546）	汉堡王，红龙虾，麦当劳
摩天大楼（121）	帝国大厦，西尔斯大厦，迪拜塔
剧场（632）	地铁，太平洋c将，标准
网络浏览器（232）	IE，火狐，Safari浏览器

表5：20个基准概念和它们的典型实例

无法解析给出增长，但受限Γ。 在第二轮中，我们可以利用更全面Γ从这些先前未句子中提取更多的对。

图10进一步示出的平均精度 ***ISA*** **每轮itera-和灰后的40个基准概念提取对。精度为迭**代继续小幅97.3％下降。对于减少的主要原因是，由于itera-和灰的进行，虽然正确对中**数Γ增加，所以确实的数目** ***不正确*** **那些。这可能会导致在超级概念检测问题。例如，考虑句子"**.....**比狗等非动物笼** **如** ***猫，狼，音响SH***

。。。**"。这可能是在一个点Γ既包含了对（狗，猫）和（非笼养动物，狗）。展望未来，随**着国语 -

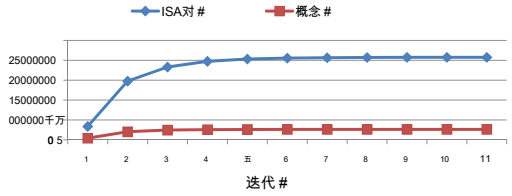


图9：数 ***ISA*** **对和概念提取**

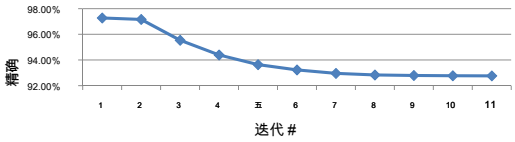


图10：精密 ***ISA*** **对提取**

的（狗，猫）增加quency，（非笼养动物，猫）的频率可能会因"非笼动物"的语料中很**少发生维持不变。因此，可能的是，似然比** ***R（狗，非笼动物）*** **超过thresh-老了一段时**间后，再算法将错误"狗"为正确的超级概念。因此，其它不正确的对像（狗，狼）和**（狗，音响SH）将推出Γ，这进一步降低了精度。**

5.3应用

典型性也已经利用其在最近的几次启用Probase-应用，如 ***分类关键字搜索[*** **9]** ***SE-mantic网页搜索[*** **39]，** ***短文本的理解[*** **34，38]，和** ***非derstanding网络表[*** **37。报告的结果表**明，在4.2节建模的典型性可以帮助前[39，37]没有触及任一地址祁门功夫，功夫坦问题，或增强现有文本的理解任务的方法性能（例如，[34]）。下面我们简要地Ÿ描述两种这**些应用程序：** ***语义网络搜索*** **和** ***短文本的理解。*** **它们说明typical-，两者均是如何在两种在第1，即概括的概念化任务中使用，** ***实例*** **和** ***抽象，*** **分别。**

语义Web搜索。 **如在第5.2节中所示，的网络搜索80％以上含有的概念和/或可以在**Probase中找到实例。这给Probase一个很好的优势，在preting用户间的意图。请看下面**的搜索查询：1）** ***ACM研究员语义Web上工作;*** **和ii）** ***数据库会议在亚洲城市。*** **这些查询**的用户意图是明确的。然而，目前的基于关键字的搜索引擎无法提供良好的结果，因为它们返回的页面与准确，字对字相匹配的短语，如"ACM研究员"，"数据库会议"和"亚洲城市"。

我们建立了一个新的语义搜索原型来处理这些语义查询[39]。由于在Probase概念的大覆盖范围，我们可以很容易地识别用户查询中的概念。然后，我们通过典型性的得分**与他们最典型的事例是代的概念重写查询。例如，我们可以更换** ***数据库会议*** **同** ***SIGMOD，VLDB，*** **等，并更换** ***亚洲城市*** **同** ***北京，新加坡，东京，*** **等重新连接奈德查询成为** ***SIGMOD在北京，*** **等等。因为有两个实例的许多可能的组合的两个概念，我们用** ***词汇联想[*** **39]实例和**关键字之间，以确定最佳一对用于取代实例。我们evalu-通货膨胀表明，平均而言，从这些改写查询返回的结果中的约80％是由用户认为与此相关，

与搜索查询原在谷歌或Bing的结果低于50％相比。

短文本解读。 **了解简短文本（例如网页搜索，微博，锚文本）是很重要的许多应用程序。STA-tistical方法，如 *主题模型[*3]把文字作为 *文字包*，和发现 *潜在主题* 从文本。然而，网络nding潜在主题并不等于理解。甲潜在主题是由一组字表示，由没有 *明确的* 概念**。更进一步，袋的词的方法通常不用于不具有足够的信号短文工作。

Probase使机器通过基于典型性进行Bayesian分析从一组单词的概念化 *T（ X | I）*。

例如，给定一个字 *印度*，机器可以推断其最 *典型* 的概念，如 *国家* 要么 *区域*。鉴于这两个词，*在 - 直径* 和 *中国*，最典型的概念成为 *亚洲国家* *要么 发展中国家*，等加一个字，*巴西*，顶部的概念可能会成为 *金砖四国* 要么 *新兴市场*，等。在重新工作迟来[34]，我们概括本概念化方法一般 *条款（* 不是说说而已是实例）。然后，我们基于由Probase提供概念上的信号（和它们的典型性分数）群集Twitter消息。我们表示每个鸣叫为具有其最典型的概念作为特征的矢量，并且执行K-均值聚类。结果胜过所有现有的方法，如LDA [3]。

六，结论

在本文中，我们提出了一个框架，在 - 自动的ferences开放的领域，从整个网络的概率分类。这种分类，给我们所知的，是目前国内规模最大的中包含的概念数量方面最全面的。它的概率模型允许既精确又暧昧的知识整合，甚至容忍不一致性的，哪些是共同在网络上的错误。更多]。更重要，这种模式使概念和实例，这将获益科幻吨广泛的需要文本的理解应用程序之间的概率推理。

参考文献：

[1] S. Auer小，C. Bizer，G. Kobilarov，J.莱曼，R. Cyganiak，和
Z.艾夫斯。DBpedia中：一个开放的数据网络核心。在 *ISWC / ASWC*，
722-735页2007年，。

[2] M.万古，MJ Cafarella，S. Soderland，M.布罗德黑德，和
O.奥尼。从网上公开信息的提取。在 *IJCAI*，
2670年至2676年2007页。

[3] D. Blei和J. Lafferty的。主题模型。在 *文本挖掘：CLASSI科幻阳离子*，*集群和应用。* 泰勒＆弗朗西斯，2009. [4] P.布卢姆。胶水的精神世界。 *性质*，421：212-21
3，2003年一月[5] K. Bollacker，C.埃文斯，P. Paritosh，T.斯特奇，和J.泰勒。

游离碱：为构建人类知识协同创建图形数据库。在 *SIGMOD*，2008年。

[6] SA卡拉巴洛。上位词标记名词的自动化建设
从层次文本。在 *ACL*，1999年。

[7] A.卡尔森，J. Betteridge，B. Kisiel的，B.落地，ERH Jr.和TM
米切尔。向为永无止境的学习语言的架构。在 *AAAI*，2010。

[8] P. Cimiano，A. Pivk，LS Thieme，第和S.施塔布。学习分类
从证据的异构数据源的关系。在 *在2004年ECAI本体学习和人口研讨会论文集*，2004年。

[9] B.丁，王H.，R.晋，汉J.和Z.王。优化指数
对于分类关键字搜索。在 *SIGMOD*，2012。

[10] D.尼，M.布罗德黑德和O.奥尼。命名复杂的定位
在网络文本中的实体。在 *IJCAI*，2733至2739年页，2007。[11] D.尼，O.奥尼和S. Sode
rland。的概率模型
冗余信息提取。在 *IJCAI*，2005年。

[12] O.伊兹欧尼，MJ Cafarella，D.唐尼，S.角，A.-M. 波佩斯库，
T.振荡，S. Soderland，DS虚焊，A.耶茨。在knowitall网络规模的信息提取。在 *万维网*，1
00-110页，2004。[13] C. Fellbaum，编辑器。 *共发现：电子词汇数据库。* MIT

出版社，1998年。

[14] M.弗雷什曼。命名实体的自动次范畴。在
***ACL（姐妹篇）* 25-30页，2001。[15] M.弗雷什曼和EH Hovy。命名细粒度CLASSI网络阳离子**
实体。在 *COLING*，2002年。

[16] MA赫斯特。上下义词的自动采集，从大文本
语料库。在 *COLING*，539-545页，1992。[17] T.李，Z.王，H. Wang和S.黄。网络规模分类
清洗。在 *VLDB*，2011。

[18] DB Lenat和RV哈。 *建设大型知识为基础的*
系统：表示与推理中的Cyc项目。
Addison-Wesley出版社，1989年。

[19] P.李，王H.，H.李和吴X.。稀疏信息提取
基于语义上下文。在提交，2012年[20]李正东，H.李，王H.，和X.周。克服语义漂移

网络规模的信息提取。下提交，2012。[21] C. Matuszek写，MJ Witbrock，RC Kahlert，J.卡布拉尔，D.施耐德，
P. Shah和DB Lenat。搜索常识：从网络中填充CYC。在 *AAAI*，1430年至1435年页，200
5。[22] G.墨菲。 *大书的概念。* 麻省理工学院出版社，2004 [23] D.纳多和S.关根。命名实体识别的调查和

CLASSI科幻阳离子。 *Lingvisticae Investigationes*，30（1）：3-26，2007。[24] R.纳维利。多义：调查。 *ACM COMPUT。*
***监测网*，41（2），2009年。**

[25] M.帕斯卡。组织和搜索事实的万维网 -
步骤二：利用群众的智慧。在 *万维网*，2007年。

[26] SP Ponzetto和M. Strube。从推导一个大型的分类
维基百科。在 *AAAI*，2007年。

[27] A.特尔，S. Soderland和O.奥尼。这是什么，反正：
自动上位词发现。在 *AAAI春季研讨会学习通过阅读和学习阅读*，2009年。

[28] E.西格尔，D.科勒和D. Ormoneit。概率抽象
层次结构。在 *NIPS*，913-920页，2001。[29] B.邵，王H.，和Y.李。三位一体图形引擎。技术
报告显示，微软研究院，2012年[30] B.邵，王H.，和Y小。管理和挖掘大图：

系统和实现。在 *SIGMOD*，2012。

[31] P.辛格，T.林，E.米勒，G.林，T帕金斯和W.李竹。打开
心灵常识：普通大众知识获取。 *在移动到有意义的互联网系统：CoopIS，DOA和ODBASE*，1223至1237年页，2002。[32] R.雪，D. Jurafsky，和伍AY。语义分类归纳

从异质性的证据。在 *ACL*，2006年。

[33] R.雪，S.普拉卡什，D. Jurafsky，和伍AY。学习合并
词义。在 *EMNLP-CoNLL*，1005-1014页，2007。[34]宋Y.，H.王，王Z.，H.李，和W.陈。短文本
概念化使用概率知识库。在 *IJCAI*，
2011。

[35] FM苏查内克，G. Kasneci，和G. Weikum。药园：一个核心
语义知识。在 *万维网*，697-706页，2007。[36] C.托马斯，P.梅拉，R.布鲁克斯和AP谢斯。越来越科幻视场
兴趣 - 使用域模型提取的扩大和减少战略。在 *Web Intelligence中*，496-502页，2008。[37] J.王，王Z.，H. Wang和KQ朱。理解上表

网页。技术报告，微软研究院，2010。[38]王S.，Y.宋，王H.和Z张。在很短的理解

文本。在提交2012年。

[39]王Y.，H.李，王H.，和KQ朱。朝在主题搜索
网页。技术报告，微软研究院，2010。[40] W.吴，李H.，H. Wang和KQ朱。Probase：一种概率
分类文本的理解。技术报告，微软研究院，2012。