arXiv:2303.04532v1 [cs.IR] 8 Mar 2023

将类别基数比较视为费米问题

Shrestha Ghosh 马克斯·普朗克信息学研究所 萨尔兰大学 萨尔布吕肯，德国
ghoshs@mpi-inf.mpg.de

Simon Razniewski 马克斯·普朗克信息学研究所 萨尔布吕肯，德国 srazniew@mpi-inf.mpg.de

Gerhard Weikum 马克斯·普朗克信息学研究所 萨尔布吕肯，德国 weikum@mpi-inf.mpg.de

摘要

类别基数比较问题很难回答，也带来了独特的挑战。此类问题需要一定程度的推理，因为网页文档和知识库这类不可或缺的信息来源，很少会直接存储问题的答案，例如“宇航员多，还是诺贝尔物理学奖得主多？”我们通过利用三个来源获取类别的绝对基数以及正交子组的基数，来解决类别基数比较问题。我们提出了用于聚合部分覆盖信号的新技术，以获得更可靠的估计，并在包含 4005 个类别对的数据集上对其进行评估，准确率达到 83.7%。

关键词

实体类别、比较型问题、类别基数估计

ACM参考文献格式：

Shrestha Ghosh、Simon Razniewski 和 Gerhard Weikum。2023年。类基数比较：一个费米问题。发表于（已录用，收录于 *WWW'23*）。ACM，纽约，纽约州，美国，5页。https://doi.org/10.1145/nnnnnnn.n

1 引言

动机与问题。宇航员多，还是诺贝尔物理学奖得主多？核电站多，还是天主教大教堂多？湖泊多，还是河流多？地球上的机场比轨道上的卫星多，还是反过来？这类比较型问题激发了我们的好奇心，却往往出人意料地难以回答。对于某些比较，有权威的官方来源提供可靠（尽管未必是最新的）数据。然而，在大多数情况下，人们自然会求助于在线信息源，例如知识库（KB，例如维基数据）、搜索引擎（SE，例如必应）的结果，或大语言模型（LM，例如 GPT）。不过，这些来源在收录哪些内容、遗漏哪些内容方面都存在偏见，而且往往会提供误导性信号，导致人们得出错误结论。

例如，维基数据知识库显示，河流超过 40 万条，城堡约有 23,000 座。虽然我们发现河流的数量接近 300,000 [9],，但城堡的数量相当难以确定；仅欧洲的城堡数量就可能在 40 万至 130 万座之间[1]。

在以不同措辞的查询进行探测时，主要搜索引擎得出的河流数量为 250,000 条（美国），城堡数量为 10,000 座（中世纪欧洲），这表明河流远多于城堡，与现实相悖。甚至连在问答任务中表现出色的 GPT-3 语言模型也完全答错了。给它三个示例问题，随后提出河流数量的问题，它回答“地球上估计有 *130* 万条河流”；而对于城堡数量，它回答：“世界上大约有 *900* 座城堡。”显然，这两个主要来源都不能如实反映现实；在线世界本质上受限于信息不完整和偏见。

另一方面，聪明的人有时能够审慎地选择在线信息源作为线索，再以巧妙的方式结合多种线索，最终对类别基数（至少达到正确的数量级）以及类别之间的相对比较（只需判断哪个更大）得出合理估计。恩里科·费米是20世纪初的一位诺贝尔物理学奖得主，以擅长此类估算而闻名；因此，这类问题也被称为费米问题 [26]。本文旨在模拟聪明人的方法。我们介绍并研究了各种可供利用的在线信号，以深入了解它们在不同实体类别领域中的优势和局限（例如职业、创意作品与人造实体

nnnnnn

对象）。此外，我们提出了新颖的技术，将覆盖范围不完整的信号聚合起来，从而得到更可靠的估计，判断给定的两个类别中哪一个在现实世界中拥有更多实例。

方法与贡献。本文聚焦于优势估计：两个类别中哪一个的基数更大。我们从三个来源获取基数的数值线索：通过 SPARQL 查询访问维基数据知识库 [24], 使用 CoQEx 方法并通过精心设计的查询访问必应搜索引擎 [7], 以及使用各种提示词的 GPT-3 语言模型 [2]。这些来源给出的绝对基数往往完全错误，因此我们仅将其视为用于进一步推断的信号。

缓解这些偏见影响的关键思路是进一步考察类别的子组，例如按国家或地理区域（如北美洲、东亚等）划分的演员或机场。这些子组与正在比较的类别彼此正交。我们研究的假设是，类别子组的估计值能够提供更可靠的基数估计，至少对某些子组以及相对比较而言如此。随后，本文提出的不同技术会对这组更大、更细粒度的信号进行聚合。

评估数据集由来自 6 个不同领域的 4005 个类别对组成。我们的主要发现是，聚合信号这一新颖技术能够显著改善优势估计，相较于直接来源信号，其准确率超过 80%。

2 相关工作

我们重点关注数字世界中的偏见，以及信息来源的完整性，因为这两者对优势估计的影响最大。

[1]https://www.quora.com/How-many-castles-are-there-in-Europe/answer/Michael-Burke-339; https://www.dw.com/en/does-germany-really-have-25000-castles/a-42350502

录用于 *WWW'23*，奥斯汀，2023年。ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

数字世界中的偏见。维基百科是通用知识的重要来源，被用于知识库的自动构建，但存在隐性偏见和显性偏见 [10]，以及性别偏见等特定的人口统计学偏见 [22]。Common Crawl [3] 是另一个规模庞大的文本信息数据集，容量约为维基百科（约 5.6 TB）的 50 倍，也远大于 BookCorpus（6 GB）。在用于大型语言模型的预训练之前，它需要经过大量处理 [18]。一项案例研究表明，经过筛选的 C4 数据集对少数群体造成了不成比例的影响 [4]。另一项研究探讨了网络本身存在的各种偏见及其潜在影响 [1]。由于多种原因，知识库中的表征并不均衡，其中包括但不限于报告偏差 [8]，以及数据偏差、模式偏差和推断偏差 [11, 21]。

信息来源。众所周知，通用知识库并不完整 [20, 25]，即使其覆盖范围随着时间不断扩大 [19]。生物学中的物种估计技术已被用于估算基数 [15, 23]。标记重捕法等技术假设样本取自现实世界；将其应用于知识库编辑历史会导致严重低估，例如预测人类大约有 400 万 [15][2]。如今，强大的搜索引擎能够从其内部知识库中提供结构化答案，并在顶部摘要中突出显示最可能的答案。即便如此，对于更复杂或不太热门的问题，结构化答案仍不常见。这种较低的召回率促使研究人员探索如何根据多个搜索引擎摘要回答实体计数问题 [6, 7]，以及如何从网络中大规模挖掘数量信息 [5]。研究表明，语言模型在回忆事实信息方面效果显著 [13, 17]。然而，它们缺乏可审查性，依赖高质量的提示词 [12],，而且已知在数值/计数信息方面表现不佳 [14]。

3 类别基数比较

让我们继续探讨这个问题：河流比城堡多吗？原则上，这类问题可以分解为基数问题：有多少条河流？以及有多少座城堡？，然后比较两者的答案。因此，我们确定了三个与比较问题相关的问题陈述，它们的信息量和难度呈反向关系。

1. 基数估计：类别河流和城堡的绝对基数是多少？ 2. 比例估计：它们的比率是多少？ 3. 优势估计：河流和城堡哪个更多？

尽管基数估计提供的信息最为丰富，但它也是最具挑战性的任务。估计方法可能会给出不切实际的结果，与真实值相差几个数量级；而对于某些类别，甚至可能不存在广泛认可的真实值。

考察类别对之间的关系可以减轻不确定性的影响：即使城堡的数量没有广泛认可的统计值，大多数估计也可能一致认为城堡比河流多。确定实际比例固然可取，但这同样受到不确定性的影响。

本文聚焦于最容易入手的优势估计任务：确定一个类别是否比另一个类别大。直观而言，作为人类，我们可以从观察中推断出城堡比河流多，因为沿着一条河通常会有好几座城堡。然而，机器无法进行这样的推理。当类别不具备这类常见观察依据时，例如比较河流和卫星的数量，这项任务即使对人类来说也很困难。尽管如此，仍有知识库实体和网页文档中的实际计数等证据可供利用，以预测这两个类别中哪一个更大。我们识别出用于预测一个类别是否大于A B的基数信号，并将输出变量定义为：

$$O_{|A|>|B|} = \begin{cases} 1 \text{ 如果预测值大于 } |B|, \ -1 \\ \text{如果预测值小于 } |B|, \text{ 若预测器弃权则为 } 0。 \end{cases} \quad (1)$$

3.1 基数信号基础

我们从三个不同来源获取两种类型的基数信号。

信号来源。我们使用三个不同的信号来源：知识库、搜索引擎和语言模型。每种来源都提供了不同的视角：知识库中涵盖的实体、网络上的热门内容，以及语言模型提炼出的信息。具体而言，我们考察维基数据 [24], 必应的前 50 条搜索结果 [16] 以及 GPT-3 [2]。

1. **知识库**。在这里，我们编写 SPARQL 查询，从维基数据中检索每个类别的实体数量，并优先采用人工标注的查询，而非知识库问答系统。由于现有系统难以编写出最准确的 SPARQL 查询，知识库问答系统的表现不如人工标注的查询。.
2. **搜索引擎（SE）**。使用 CoQEx 系统，根据排名前 50 的搜索引擎结果摘要 [6] 推断数量，并返回最有把握的类别数量。
3. **语言模型（LM）**。在少样本设置下，将 n 个数量问题及其答案作为提示词，检索 $(n+1)^{th}$ 问题中类别的数量。具体而言，我们向 GPT-3 模型提供三个示例以及目标问题。对于所有查询，示例和模型参数均保持不变。我们使用 Python Quantulum3 库[3]处理文本输出，以提取类别的数量。

信号类型。我们考虑两种类型的信号。

1. **根信号**。表示目标类别自身数量的信号，例如河流（全球范围内）。2. **子组信号**。表示目标群体的子组数量的信号，例如德国的城堡。

什么是正交子组？如何选择它们？

原则上，可以针对一系列子组（例如，195个国家、按年份/十年期、按状态等）计算子组信号。下文我们重点关注二十国集团国家 [27],，因为关于这些国家，来源往往提供更可靠的信息。

还要注意，子组信号通常不会相加得出根信号，这既有实际原因（某些子组的数据不可用或不完整），也有原则性原因（某个实体可能属于多个子组，也可能不属于任何子组）。

[2] https://cardinal.exascale.info/

[3] https://pypi.org/project/quantulum3/

3.2 信号聚合

除了直接使用根信号，我们还会汇总基本信号来预测类别比较。我们建议分三个层次进行：首先汇总子组信号，然后纳入根信号，最后按来源汇总所得信号。

I. 子组聚合。。

1 按多数规则：$O_{M}^{|A|>|B|}$ 为 1，若 A 的子组中至少有 θ_M % 的子组所含实体多于 B 对应子组，则为 -1，若至少有 θ_M 的子组所含实体少于 B 对应子组；否则为 0。.

2 按显著性：我们进行单侧 t 检验；当 A 的子组基数均值大于 B 时，检验 A 是否大于 B。当相反情况成立时，则检验 A 是否小于 B。在前一种情况下，若 p 值 $\leq \alpha$，则 A 的子组分布显著大于 B，且 $O_{S}^{|A|>|B|}$ 为 1。若后一种情况下 p 值 $\leq \alpha$，则结论相反，且 $O_{S}^{|A|>|B|}$ 为 -1。若任一情况下 p 值 $> \alpha$，则预测为 0。

II. 根信号与子组聚合。在此，最终预测是以下预测的平均值：i) 比较根信号所得的预测，ii) 对子组信号进行多数投票所得的预测，以及 iii) 对子组信号进行显著性检验所得的预测。我们为多数投票和显著性预测赋予权重 $W \in [0,1]$，使得，

$$Ensemble_{|A|>|B|}^{source} = \frac{1}{3}(O_{|A|>|B|}^{source} + W_M \cdot O_{|A|>|B|}^{source,M} + W_S \cdot O_{|A|>|B|}^{source,S}) \quad (2)$$

因此，当多数结果为 A 大于 B 时，权重 W_M > 即为多数比例；当 θ_1 B 较大时，则为（即 多数比例）$W_M = 1 -$。显著性预测的权重为 (1-p 值)，因此 p 值越低，显著性越高。

III. 最终来源聚合。在最后一级聚合中，我们通过以下方式合并所有来源的预测：

多数投票：如果任意两个来源的结果一致，则选择该标签；否则，预测器弃权。

加权投票：对于每种信号聚合情况，我们都会训练一个逻辑回归分类器，以学习各来源的权重。

$$ln(p(|A| > |B|)) = -ln(1 + e^{-\sum_{source} W_{source} \cdot Ensemble_{|A|>|B|}^{source}}) \quad (3)$$

$$Ensemble_{|A|>|B|}^{weighted} = \begin{cases} 1 & \text{if p(|A|>|B|) > p(|A|<|B|),} \\ -1 & \text{if p(|A|>|B|) < p(|A|<|B|),} \\ 0 & \text{otherwise.} \end{cases} \quad (4)$$

4 实验

数据集。我们构建了一个包含 90 个类别、覆盖 6 个领域的真实标注（GT）数据集（表 1）[4]。我们的评估集包含 $\binom{90}{2} = 4005$ 个类别对组合，其中 $6 \times \binom{15}{2} = 630$ 个为域内类别对，即两个类别属于同一领域；其余 3375 个类别对为跨域类别对，即两个类别属于不同领域。为评估任务的难度，我们计算所有类别对真实基数之比的数量级。我们认为，基数接近的类别对比基数相差多个数量级的类别对更难预测。数据集中基数接近的类别对更多，而基数之比超过 10^4 个数量级的类别对不到 8%，

[4]数据集链接：https://github.com/ghoshs/class_cardinality_comparison。

表 1：领域及示例类别。

领域	#类别数	示例
创意作品	15	电影、棋盘游戏、书籍
地理实体	15	湖泊、城堡、大坝
人造物体	15	人造卫星、潜艇
职业	15	政治家、演员、物理学家
组织	15	大学、足球俱乐部
物种	15	蛇、昆虫、鱼

类别对的基数之比超过 10^4 个数量级的不到 8%。例如，航空公司（5,000）与国家公园（3,369）的基数之比为 1.4 : 1（较大值：较小值），而航空公司与政治家（6,500,000）的基数之比为 1.3×10^3 : 1。

指标。我们主要通过每个信号的准确率来衡量其性能，即正确预测占所有样本的百分比。为了进一步判断准确率低是因为经常弃答还是预测错误，我们还报告弃答率和精确率；精确率是正确预测占未弃答样本的百分比。

参数。我们使用 GPT-3 的*text-curie-001*模型，将温度设为 0，并将生成时的最大令牌数设为 15。子组聚合参数为 θ_M 和 α，我们分别将它们设为 0.5（多数超过 50%）和 0.05。

训练逻辑回归分类器时，我们将数据集划分为训练集和测试集（80:20）。我们在训练数据上进行 5 折交叉验证，以确定正则化超参数。我们报告在包含 4005 个类别对的整个数据集上的评估结果，但加权投票集成除外；该集成在包含 800 个样本的测试集上进行评估。

人类基线。我们抽取了 100 个类别对，其中 50 个为域内类别对，50 个为跨域类别对，以评估非专家人类表现，模拟可能并非领域专家的普通用户。每个类别对由三名 MTurk 标注员进行标注，他们会看到每个类别的简要描述，并回答以下问题：

(1) 现实生活中哪个类别的实体更多？
类别 *1*、类别 *2*、数量相同、无法确定

(2) 你有多确定？非常确定、估计、猜测、毫无头绪此外，标注员可以在自由文本栏中说明作答理由。该任务分为两种设置：(i) 闭卷设置，标注员应在不查阅外部来源的情况下作答；(ii) 开卷设置，鼓励标注员进行网络检索。在闭卷设置下，每道题限时 3 分钟；在开卷设置下，每道题限时 8 分钟。

两种设置下的准确率相近（75% 对 76%），但精确率的提升幅度更大（从 79% 提升至 85%）。也就是说，在开卷设置下，额外的证据使标注员更常选择弃答，而不是猜测。

最先进的基线。信号聚合方法与来自三个来源（知识库、搜索引擎、语言模型）的根信号的最先进的基线进行比较。

结果 1：按来源划分的性能。表2展示了各来源在不同信号下的准确率。在三个来源中，只有语言模型的表现超过了人类基线。我们发现，对根信号和聚合的子组信号进行集成后，三个来源的表现都很好。聚合子组信号和根信号可将

表2：基数信号的准确率（%）。

来源	根信号（1）	子群聚合		根节点和子群信号集成		
		多数（2）	T检验（3）	(1)+(2)	(1)+(3)	(1)+(2)+(3)
KB	64.7	57.0	36.0	61.5	**65.9**	61.8
SE	65.4	67.2	39.6	65.7	65.4	**68.4**
LM	74.4	75.8	57.3	77.1	75.8	**79.4**
跨来源集成（知识库、搜索引擎、语言模型）						
多数投票	77.8	76.3	42.9	76.8	78.7	**78.9**
加权投票	78.2	76.2	81.2	79.3	**83.7**	81.3
非专家人类基线						
闭卷	75.0					
开卷	76.0					

表 3：按领域划分的根信号与子组聚合准确率（%）。

领域	KB	SE	LM	Best
创意作品	62.8	54.2	**83.8**	LM
地理实体	**77.1**	60.9	70.4	KB
人造物体	26.6	77.1	**96.1**	LM
职业	57.1	**80.0**	74.2	SE
组织	58.0	72.3	**88.5**	LM
物种	61.9	**78.0**	63.8	SE
跨领域	62.6	68.0	**79.4**	LM
All	61.8	68.4	**79.4**	LM

表 4：数据集中的示例，按难度从低到高排序。

类别 1	类别 2	GT 比率	根信号 知识库	搜索引擎	语言模型	基本信号集合* 知识库	搜索引擎	LM	源集合** 多数	加权	备注
网站	宗教文本	$6\times10^6:1$	✓	✓	✓	✓	✓	✓	✓	✓	来自所有来源的强信号。
细菌物种	蜜蜂物种	$3.15\times10^5:1$	✓	x	x	✓	x	x	x	x	知识库根信号比为 4.8:1，可得出正确预测。
书籍	绘画	$1.3\times10^3:1$	x	x	✓	x	x	✓	x	✓	加权投票从语言模型中捕捉到弱信号。
学校教师	医院	$6.07\times10^2:1$	x	✓	x	x	✓	✓	✓	✓	来自搜索引擎和语言模型中子组聚合的正确信号。
城市	岛屿	9.7 : 1	x	✓	x	x	✓	✓	✓	x	加权投票因错误的知识库信号过强、正确的语言模型信号较弱而失效 +
参与者	架构师	1。48 : 1	✓	x	x	✓	x	x	x	✓	加权投票利用了强大的知识库信号。

* 知识库的最佳集成：*(1)+(3)*；搜索引擎和语言模型的最佳集成：*(1)+(2)+(3)* 基于基本信号：根信号 (1)、多数聚合 (2) 和 T 检验聚合 (3)。 ** 多数投票的最佳来源集成：*(1)+(2)+(3)*；加权投票：*(1)+(3)*。

弃答率从 0.5% 降至 5%，并在各来源中始终低于 1%。进一步检查后我们发现，t 检验子组聚合的准确率最低，尽管其精确率较高（语言模型：87%，搜索引擎：83%），这是由于弃答率极高（语言模型：34%，搜索引擎：52%）。

结果 2：来源聚合。从表2中可以看出，使用监督学习来学习权重，能够提高不同聚合策略的稳健性，因为准确率不会低于 76%，因此表现优于任何单一来源。相反，针对来源进行多数投票，仅在根信号上最有效。值得注意的是，在加权投票策略中，按 T 检验进行的子组聚合表现出色：仅聚合 T 检验信号时准确率达到 81%，聚合根信号和 T 检验信号时则达到 83.7%。在这两种情况下，我们都发现语言模型的系数较高，其次是搜索引擎，知识库的系数最低。

结果 3：按领域划分的表现。我们在表 3 中，针对根信号和子组信号集成，分析了 6 个领域中各自的类别对。请注意，没有任何一个来源始终胜出；若按领域采用表现最佳的来源，准确率可超过 77%。自行车或智能手机等日常物品的估算值很容易在网络上获取，这一点也反映在搜索引擎（SE）和语言模型（LM）的准确率上，分别为 77% 和 96%。知识库（KB）在地理实体方面的准确率最高（77.14%），大幅优于另外两个来源。

讨论。对于当前的搜索引擎和语言模型而言，直接比较类问题更具挑战性。例如，对于问题河流和湖泊哪个更多？，必应给出的答案文不对题，而 GPT-3 则给出了明确答案，称湖泊更多。当被要求解释时，答案却反了过来。表 4 展示了我们数据集中的一些示例，以及表2中每一行的根信号预测和表现最佳的集成预测。

5 结论

在这项工作中，我们利用来自不同信息来源的信号，对现实世界中的两个类别进行优势估计。我们提出了通过在正交子组内以及多个来源之间聚合来组合基数信号的方法。实验结果表明，基于子组聚合、跨来源构建的集成在优势估计方面表现出色。

参考文献

[1] Ricardo Baeza-Yates。2018年。网络上的偏见。载于*ACM*通讯。[2] Tom Brown 等。2020年。语言模型是少样本学习者。载于*NeurIPS*。[3] Common Crawl。2008年。Common Crawl 数据集。https://commoncrawl.org/。[4] Jesse Dodge 等。2021年。大型 Webtext 语料库文档记录：对 Colossal Clean Crawled Corpus 的案例研究。载于*EMNLP*。[5] Yanai Elazar 等。2019年。狮子有多大？推导定量属性的分布。载于*ACL*。[6] Shrestha Ghosh、Simon Razniewski 和 Gerhard Weikum。2022年。利用解释性证据回答计数查询。载于*SIGIR*。[7] Shrestha Ghosh、Simon Razniewski 和 Gerhard Weikum。2022年。利用文本中的结构化答案回答计数问题。载于*JoWS*。[8] Jonathan Gordon 和 Benjamin Van Durme。2013年。报告偏差与知识获取。载于*AKBC*。[9] Günther Grill 等。2019年。绘制世界上自由流动的河流。载于《自然》。[10] Christoph Hube。2017年。维基百科中的偏见。载于*WWW Companion*。[11] Krzysztof Janowicz 等。2018年。消除知识图谱中的偏见：为什么女性总统不同于女性教皇。载于*ISWC* 工作坊。[12] Zhengbao Jiang、Frank F. Xu、J. Araki 和 Graham Neubig。2020年。我们如何知道语言模型了解什么？载于《*ACL* 汇刊》。[13] Vladimir Karpukhin 等。2020年。用于开放领域问答的密集段落检索。载于*EMNLP*。[14] Bill Yuchen Lin 等。2020年。鸟有四条腿？！NumerSense：探测预训练语言模型的数值常识知识。载于*EMNLP*。[15] Michael Luggen 等。2019年。协作式知识图谱的非参数类完整性估计器——以维基数据为例。载于*ISWC*。[16] 微软。2022年。必应搜索 API。https://www.microsoft.com/en-us/bing/apis/bing-custom-searc

h

[17] Fabio Petroni 等人。2019年。语言模型能否作为知识库？发表于*EMNLP-IJCNLP*。[18] Colin Raffel 等人。2020年。利用统一的文本到文本转换器探索迁移学习的极限。发表于*JMLR*。[19] Simon Razniewski 和 Priyanka Das。2020年。结构化知识：我们取得进展了吗？对 19 年间知识库覆盖率的外部研究。发表于*CIKM*。

[20] Simon Razniewski、Fabian M. Suchanek 和 Werner Nutt。2016年。但我们实际上知道什么？发表于*AKBC*。[21] Tara Safavi 等人。2021年。关于首届 AKBC 2020 自动知识图谱构建中的偏见研讨会的报告。发表于*SIGIR Forum*。[22] Jiao Sun 和 Nanyun Peng。2021年。男人当选，女人结婚：维基百科中的事件性别偏见。发表于*ACL-IJCNLP*。[23] Beth Trushkowsky、Tim Kraska、Michael J. Franklin 和 Purnamrita Sarkar。2013年。众包枚举查询。发表于*ICDE*。[24] Denny Vrandečić 和 Markus Krötzsch。2014年。维基数据：一个免费的协作式知识库。载于*ACM*通讯。[25] Gerhard Weikum 等。2021年。机器知识：全面知识库的创建与整理。载于*Foundations and Trends Databases*。[26] 维基百科。2022年。费米问题。en.wikipedia.org/wiki/Fermi_problem。[27] 维基百科。2023年。G20国家。https://en.wikipedia.org/wiki/G20。