挖掘网络预测未来事件

基拉Radinsky
Technion-Israel Institute of Technology,
以色列海法
kirar@cs.technion.ac.il

Eric Horvitz
微软研究院
美国华盛顿州雷德蒙德
horvitz@microsoft.com

摘要。

我们描述并评估了从包含22年新闻故事的语料库中学习预测即将发生的感兴趣事件的方法。我们考虑了在世界上这些事件发生之前识别疾病爆发、死亡和骚乱可能性显著增加的例子。我们提供了方法和研究的细节，包括从新闻语料库和多个网络资源中自动提取和概括事件序列。我们评估了该方法对系统隐瞒的现实世界事件的预测能力。

类别和主题描述符

I.2.6[人工智能]:学习;I.2.1[人工智能]:应用和专家系统

一般条款

算法,实验

关键字

新闻语料库，预测未来新闻，从Web内容学习

1. 介绍

马克·吐温有句名言:“过去不会重演，但会押韵。”本着这种反思的精神，我们开发并测试了一些方法，利用从《纽约时报》(NYT)存档的22年新闻报道中捕获的大规模数字历史，对未来人类和自然事件感兴趣的可能性进行实时预测。我们描述了如何通过概括报道的新闻事件序列中的特定过渡集来学习预测未来，这些新闻事件是从1986-2008年的新闻档案中提取的。除了新闻语料库之外，我们还利用了来自免费Web资源的数据，包括Wikipedia, FreeBase, OpenCyc和GeoNames，通过领英数据平台[6]。目标是建立预测模型，从特定的事件序列集进行概括，根据近期新闻提要中观察到的证据模式，提供未来结果的可能性。我们提出这些方法，作为在世界上目标事件发生之前生成可操作预测的一种手段。

我们描述的方法在新闻提要上运行，可以提供大量的预测。我们展示了挖掘数千个新闻故事来为一系列预测问题创建分类器的预测能力。我们展示了三个预测挑战的预测示例:对即将到来的疾病爆发、死亡和骚乱的前瞻性警报。这些事件类别很有趣，可以作为预测的例子，作为关注的先驱者，指导干预措施，可能能够更好地改变结果。我们将这些方法的预测能力与几个基线进行了比较，并证明了这些领域的预测精度从70%到90%不等，召回率为30%到60%。

这项工作的贡献包括自动抽象技术，该技术将分析水平从特定实体转移到考虑更广泛的观察和事件类别。抽象通过在更高层次的本体-逻辑层次中将事件识别为更一般的证据集和结果集的成员，扩大了训练集的有效规模。例如，我们可以从特定国家(例如，安哥拉和卢旺达)的事件新闻数据中学习，以建立分类器，考虑事件在大陆(例如，非洲)或以特定人口和地质属性为特征的地区的可能性。从领英数据中提取出安哥拉和卢旺达是组成非洲的更广泛国家集合的元素的知识。

例如，学习和推理方法可用于在特定区域内提供关于即将到来的霍乱爆发可能性增加的警报。霍乱是一种快节奏的感染，每年造成10万多人死亡，未接受治疗的患者死亡率超过50%。通过及时补液治疗，死亡率降至1%以下。根据对新闻报道的监测，对未来霍乱爆发可能性的推断跳跃发出警报，有助于对注意力进行分流和制定计划。例如，推断出霍乱在特定时期爆发的可能性，可以指导在风险较高的地区分配淡水的主动设计。我们所描述的方法可能有一天会被用来继续监测不断发展的新闻故事，并提供自动化的

WSDM'13, February 4–8, 2013, Rome, Italy. Copyright 2013 ACM
978-1-4503-1869-3/13/02 ...$15.00.

提醒人们关注结果的可能性增加。这种预测可以作为现有监测和通信服务的辅助，例如世界卫生组织(WHO)全球警报和反应(GAR)系统，用于协调对突发公共卫生事件的反应[1]。在测试中，我们发现，自动预测可以在几次霍乱爆发前一周提供警报(图6)。

研究疾病传播与自然灾害之间关系的流行病学家等专家也得出了类似的结论。然而，这类研究通常数量较少，采用启发式评估，并且经常是回顾性分析，而不是旨在生成指导近期行动的预测。相比之下，计算系统具有从大量数据中学习模式的能力，可以监控众多信息源，可以随着时间的推移学习新的概率关联，并且可以继续进行实时监控，预测和警告即将到来的有关事件的可能性的增加。除了在研究中容易发现或从专家那里获得的知识之外，新的关系和结果的上下文敏感概率可以通过具有长触角的计算系统发现到历史语料库和实时提要中。例如，我们描述的方法确定了安哥拉干旱和风暴之间的关系，而干旱和风暴反过来又催化了霍乱的爆发。关于霍乱下游风险的警报本可以提前近一年发布(图1)。关注较短时间范围的人类专家可能会忽略这种长期相互作用。在追求事件之间的概率影响时，计算系统可以考虑多个时间粒度和视界。除了根据即将出现的感兴趣结果的可能性增加来提醒可操作的情况外，预测模型还可以在数据推断与专家预期相反时提供指导，从而更普遍地提供帮助。在以自动化方式考虑大量观察和反馈的基础上，识别出事件发生可能性明显低于专家预期的情况是有价值的。最后，监测有关未来事件的可能性的系统通常可以更快、更全面地获取表面上看起来不太重要的新闻故事(例如，在当地报纸上发表的关于葬礼的故事，没有达到主要标题)，但这可能为更大、更重要的故事(例如，大规模骚乱)的演变提供有价值的证据。

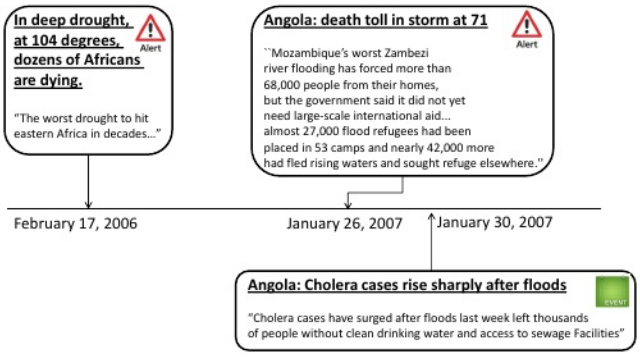

图1:安哥拉干旱后暴雨后霍乱可能性上升的例子。三角形警报图标表示几天内发生霍乱爆发的可能性显著上升的推断。

2. 事件预测

我们假设现实世界中的事件是由概率模型生成的，该模型还生成与这些事件相对应的新闻报道。我们使用新闻故事的文本to建立一个形式为$Pev_j(\tau+\Delta)ev_i(\tau)$的推理模型，用于某些未来事件$ev_j$在时间$\tau+\Delta$和过去事件$ev_i$在时间$\tau$发生(例如:今天)。例如，该模型了解到关于干旱($ev_j$)的新闻报道发生在关于洪水($ev_i$)的新闻报道之后的概率为18%。这个概率近似于这两个现实世界事件之间的关系。

给定一个目标未来事件(如霍乱爆发)，计算未来每一个可能时间的这个概率$\tau+\Delta$和每一种可能ev_i都是一个棘手的问题。我们通过关注可能因果关联的事件序列候选的一小部分来简化分析，并以这种方式定义与目标事件ev_j相关联的事件集ev_i。特别是，我们定义并从纽约时报档案新闻故事情节中提取-主题内聚有序的新闻片段集，其中包括关于单个故事的两个或多个声明性独立子句。例如，下列事件构成一个故事情节:{(非洲干旱，2006年2月17日)、(卢旺达风暴，2007年1月26日)、(卢旺达洪水，2007年1月27日)、(卢旺达霍乱爆发、01/30/2007)}。然后，我们使用这样的故事情节作为识别事件之间可能的因果关系的启发式。这个过程是通过聚类具有相似文本和语义实体的新闻故事来执行的，详见2.1节。

我们在图2中显示了该方法的组件化视图。在学习阶段的开始，系统挖掘NYT新闻语料库并提取故事情节，使用改编自知名主题跟踪和检测算法的技术[8,3,7]，将相似的文本聚类在一起(第2.1节)。接下来，我们通过关联数据项目(Section 2.2)从Web知识来源中提取信息来丰富故事情节。我们提取了各种各样的事实，包括卢旺达的人口密度、卢旺达被水覆盖的土地百分比和该国的国内生产总值等信息。我们概括了特征和事件，以增加用于构建预测模型的等效样本数量(第2.3节)。例如，我们可以从特定国家(例如，安哥拉和卢旺达)的事件数据中学习，以建立分类器，考虑更大规模(例如，更大的非洲大陆)或具有特定人口和地质特征的地区感兴趣的事件的可能性。在学习结束phase,时，系统估计概率$Pev_j(\tau+\Delta)ev_i(\tau)$，并构建一个概率分类器用于预测阶段。分类器可用于提供感兴趣事件的实时概率，例如根据先前在关于安哥拉或其概括为非洲的故事情节中获得的知识，即将发生的“安哥拉霍乱爆发”。我们构建的分类器提供了二元预测是否

1 http://www.who.int/csr/alertresponse/en/

一个事件是否会按照观察到的事件序列发生。在实验中，我们还评估目标事件的发生以及预测和发生之间的平均时间。我们展示了在实际预测事件发生前近三周提供此类警报的结果。我们将基于特定动态的感兴趣事件的确切日期预测留给未来的工作。

2.1 提取事件链

我们从NYT档案中定义和提取新闻故事情节，作为识别事件之间潜在因果关系的启发式方法。故事情节是一组具有主题凝聚力或顺序的新闻片段，其中包括关于单个故事的两个或多个声明性独立子句。例如，一条关于豺狼卡洛斯被捕的故事线包含了关于他的身份验证、他被送进监狱等等的故事。提取这类故事线的方法被称为话题检测与追踪(TDT)[8]。主题检测涉及识别故事流中的一系列关联事件。为了识别故事情节，我们修改了Inc.LM方法，这是一种主题跟踪方法，在几次比赛中被发现在这项任务中最成功[3]。假设链∈2-2$_{T\times T}$ time是所有可能的集合

故事线，其中T是所有的新闻文章，Time是时间的离散表示。我们用$t_1 <<_c t_2$表示一个由新闻文章t_1表示的事件，该事件发生在一个链c∈链中由新闻文章t_2表示的事件之前。我们使用符号t(τ)来表示由时间τ∈time的新闻故事文本t的出现所定义的事件。在因果关系只发生在故事情节中的假设下，预测的挑战是计算概率P t(τ $_i$ > τ) tj (τ)forre-{t_i|∃c∈链，$t_j <<_c t$}。

与其他用于主题检测的向量空间方法类似[7]，我们首先对具有相似文本的文档进行聚类。我们将新闻文章视为文档，并将每篇新闻文章表示为一个向量($\sigma^{1t}\cdots$)σ_n^t)，这样

$$\sigma_i^t = \mathrm{tf}_{w,t} \cdot \log \frac{|T|}{|\{t' \in T \mid w_i \in t'\}|},$$

其中|T|是所有的新闻文章，$tf_{w,t}$是文章T中单词w出现的频率。然后我们执行最近邻分析，我们使用余弦相似度度量方法为每篇文章找到最接近它的k篇文章(在我们的实验中k = 50)，定义为

$$\mathrm{sim}(t_a, t_b) = \frac{\sum_{i=1}^{N} \sigma_i^{t_a} \sigma_i^{t_b}}{\sqrt{\sum_{i=1}^{N} {\sigma_i^{t_a}}^2} \sqrt{\sum_{i=1}^{N} {\sigma_i^{t_b}}^2}},$$

在时间接近性的约束下。文章要么在一个阈值时间范围内生成，要么在链中最近一篇文章的文本中提到一篇文章的日期。我们在TDT4语料库[2]上使用时间阈值进行了几次实验，并在将链限制为14天时达到了最佳性能。这种类型的分析对于识别涵盖相同主题的文章具有很高的召回率，这是指检索到的相关实例的比例。然而，该过程的精度较低;识别出的实例中有很大一部分是误报。我们希望在保持高召回率的同时提高准确率。作为减少误码率的方法

假阳性，我们覆盖了一个偏好，即故事文章C的实体{e∈entities}的熵，定义为

$$StoryEntropy(C) = -\sum_{i=1}^{n} P(e_i \in C) \log P(e_i \in C)),$$

随着故事的发展而“缓慢”发展。类似的方法已被证明在相关的主题检测任务上提供了重大改进[2]。我们使用在异构语料库[9]上训练的条件随机场(conditional random field, CRF)来识别位置、人员和组织类型的实体。我们定义了实体计数的向量，以及从链中添加和删除文章的操作。我们使用贪婪算法，在每一步选择下一个最好的文档来添加或决定停止，如果所有剩余的文档增加的熵超过阈值α，我们从验证集评估。我们进行的实验表明，这种扩展在保持召回水平的同时提高了精度(参见第3.4节)。

我们在离线语料库上执行了后一个过程。我们注意到，研究已经解决了使用类似技术从在线新闻流中提取的问题(例如，[2])。这种在线方法可以适用于我们预测未来事件的方法。

2.2 词汇和事实特征

我们试图推断出一个预定义的未来感兴趣的新闻事件的概率，给出一个向量，表示在特定时间内发生的新闻事件。为了完成这项任务，我们为每个目标事件创建训练案例，其中每个案例都使用一组观察值或特征来表示。我们定义了词汇特征和事实特征。只有当表示未来目标事件的文本出现在链中较晚时间的文档中时，我们才将每种情况的标签设置为true。

设$w_1\cdots w_n$是表示时间τ的事件概念的单词，让一个$_1\cdots a_m$是事件概念的附加现实世界特征。我们将这些属性分别称为词汇特征和事实特征。单词w_i是使用Stanford Tokenizer从每篇新闻文章的文本中提取出来的，并使用停止词列表进行过滤。事实特征a_i是从不同的LinkedData源中提取的，特别是事件概念w_i的rdf:Property类型下的属性。例如，给定新闻标题“安哥拉:洪水后霍乱病例急剧上升”的文本，特征向量包含文本的标记化和过滤词(安哥拉、霍乱、上升、急剧、洪水)，以及描述安哥拉特征的其他特征(GDP、水覆盖、人口等)。我们将事件文本中的每个概念映射到一个LinkedData概念。如果几个概念匹配，我们基于概念文本(例如，它的维基百科文章)和新闻文章之间的相似性执行消歧，使用词袋表示。

我们表示$f_1(ev)\cdots f_{n+m}(ev)$是事件ev的特征(词汇特征或事实特征)。我们做了一个天真的简化假设，即所有特征都是独立的，并将概率$P(ev_j(\tau+\Delta)ev_i(\tau))$描述如下:

$$P(ev_j(\tau+\Delta)|ev_i(\tau)) \propto \prod_{k=1}^{n+m} P(ev_j(\tau+\Delta)|f_k(ev_i(\tau))).$$

[2] http://www.nist.gov/TDT

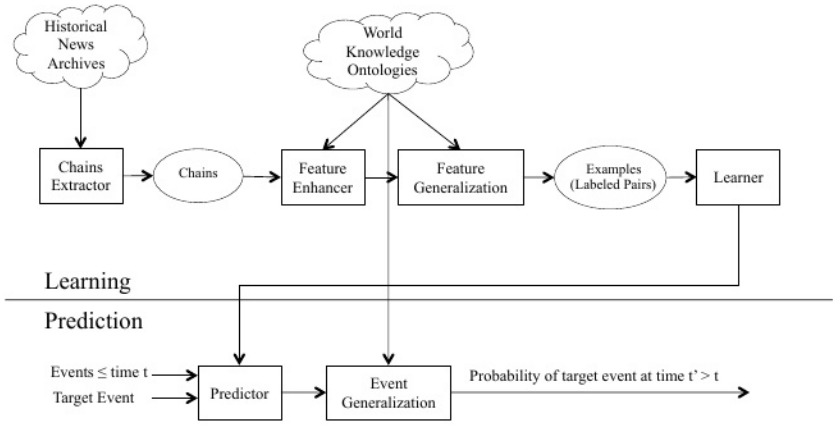

图2:事件预测管道分析的主要组成部分和流程。

使用贝叶斯规则，我们可以推导出

$$P\big(ev_j(\tau+\Delta)\big|f_k(ev_i(\tau))\big)=\frac{P\big(ev_j(\tau+\Delta),f_k(ev_i(\tau))\big)}{P(f_k(ev_i(\tau)))},$$

其中$Pev_j(\tau+\Delta)$，$f_k(ev_i(\tau))$☒is通过计算事件ev_j在同一故事线中发生的次数来从数据中评估，该事件在时间τ时具有事件ev_i的特征值f_k。类似地，通过计算具有此特征值的事件在语料库中发生的次数部分来从数据中评估$P(f_k(ev_i(\tau))$。我们基于这些学习到的概率构建一个预测器。预测器输出系统中表示的每个未来事件发生的概率。可以记录这个概率的变化，并设置阈值进行警报。

我们可能对以标量值的形式预测未来事件的各种特征感兴趣。例如，除了预测死亡与以后在视界内发生的事故或破坏相关的可能性之外，我们可能希望预测在导致死亡或与死亡相关的目标事件发生的情况下，会死亡的人数。为此，我们将死亡人数的目标预测放入一组桶中，这些桶捕获了一组相互排斥且详尽的死亡人数范围，例如，死亡人数少于10人，死亡人数大于10但小于100人，死亡人数大于100人。如果ev_j符合bin_k的关系，我们说ev_j属于bin_k。例如，对于死亡人数少于10人的箱子，如果表示ev_j的文本包含表明死亡人数的文本并且该数字小于10，则我们说$ev_j\in bin_{0-10^0}$。我们学习预测器，将事件的概率估计为belong到某个bin k, $Pev_j(\tau+\Delta)$，$ev_j(\tau+\Delta)\in bin_k ev_i(\tau)$，并输出概率最高的bin。在第3.2节中，我们提出了关于推断事件造成的死亡人数的准确性的研究结果。

2.3 学习用抽象进行预测

在我们探索的领域中，事件序列是相对稀疏的。例如，目标事件“卢旺达霍乱爆发”在新闻档案中只出现了33次。由于$Pev_j(\tau+\Delta)$，$f_k(ev_i(\tau))$☒andP($f_k(ev_i(\tau))$)这两个估计都很差，这种稀疏性可能会降低分类器的性能。在其他情况下，特征值可能不会以足够高的频率出现在数据中，从而无法推导出关于未来的高置信度推断。例如，在新闻语料库中甚至没有一次提到非洲国家科摩罗。词汇特征的这种稀疏性可能导致☒poor对概率$Pev_j(\tau+\Delta)$，$f_k(ev_i(\tau))$的估计，从而导致较差的预测器。例如，对于即将到来的大规模撤离可能性的目标预测，源自科摩罗的飓风可能是预测附近国家风暴的重要资料，这可能有助于预测这些国家的撤离情况。同样，如果系统专注于对科摩罗即将到来的风暴进行预测，即“科摩罗风暴”是目标事件，则可能没有足够的数据来评估上述概率。

我们通过采用自动抽象的过程来解决事件和特征稀疏性的共同挑战。我们不考虑只有少数历史病例的“卢旺达霍乱爆发”事件，而是考虑更一般的事件:“[非洲国家]霍乱爆发”。我们求助于网络上可获得的世界知识。一些LinkedData资源提供了分层的技术。例如，Fabian等人[24]从维基百科的内容中创建了一个isA本体。这个本体将卢旺达映射到以下概念:共和国、非洲国家、内陆国家、班图国家等。类似地，WordNet提供了将卢旺达映射到概念国家的首字母缩略词关系。

我们开发了一种自动化的方法来指导抽象。该方法决定何时对事件和特征进行泛化。由于特征是单独评估的，因此对每个特征抽象的值进行估计，以提高预测目标事件的准确性。我们使用交叉验证来评估每个特征及其抽象对训练数据的精度。我们注意到它

在不改变目标事件的情况下，不足以测量与使用抽象特征相关的精度。考虑抽象的特征[非洲国家]和目标事件“基加利之死”。在卢旺达首都基加利，由于非洲某些国家发生的事件而造成死亡的可能性很小。因此，如果某一事件发生在[非洲国家]，则该事件在首都([非洲国家])造成的死亡概率通常可能更为合适。我们现在将这种直觉形式化。

设语义网络图G是一个有边标记的图，其中每条边都是一个三元组hv_1，$_{v2}$，li和l是一个谓词(例如，“首都”)。我们寻找一条最大长度为k的路径(在我们的实验中k = 3)，它将表示抽象的概念和目标事件中描述的概念连接起来。例如，给定一个链，其中第一个事件讨论歌剧“鼻子”的大量上座率，第二个事件讨论歌剧作家德米特里·肖斯塔科维奇获得的奖项，我们在维基百科图中找到以下路径，连接$_{\text{OperasBy}}$

文章:“鼻子”- - - - - - - →德米特里·肖斯塔科维奇。后来，我们可以用类似的观察，观察歌剧《鲨鱼同志谋杀案》的大量上座率，来预测歌剧作家威廉·伯格斯马(William Bergsma)的获奖情况

$_{\text{OperasBy}}$

- - - - - - - - →威廉·伯格斯马。给定两个事件的概念c, c，由G中的节点v和v表示，我们把连接v_1和v的k的标签称为抽象路径abs (c，=l，⋯,l_k。在节点v上应用抽象，确定满足抽象路径的节点v^0，即ApplyAbs(v, abs(c_1，c_2))=v^0,)

s.t∃v_i∈v (G)(v, $_{v1}$，l_1)⋯，(v_{k-1}，v^0，l_k)∈E(G)。

当推断目标事件中每个实体en的概率和导致事件的特征时，我们使用语义层次图G^H迭代地将每个特征f抽象为更一般的概念gen(f)，计算ab-抽象路径abs(f, en)(基于语义图G)，而不是$Pev_j(\tau+\Delta)$，$f_k(ev_i(\tau))$，我们计算更抽象事件的概率，

$$P\left(ApplyAbs\left(ev_j(\tau+\Delta), abs(f_k, en)\right), gen(f_k)(ev_i(\tau))\right).$$

在泛化目标事件时，也会进行类似的过程。在这种情况下，对抽象的目标事件计算概率，对具体的目标事件计算精度。对于目标事件中的每个实体en和导致事件的一个特征，我们希望使用语义层次图G^H将每个特征f迭代地抽象为一个更一般的概念gen(f)(在我们的实验中我们使用了IsA和InCategory关系)。

图3显示了抽象过程的伪代码。给定一个目标和一个可能的因果事件，该过程的目标是估计因果事件或其任何抽象导致目标事件的概率。该算法给出几个参数作为输入:一个目标事件(例如，卢旺达霍乱爆发)，表示为目标，一个发生在时间τ(原因)的事件，系统提取的故事情节，表示为链，层次图G^H，语义图(G)，以及指定最大抽象程度的一些参数(k)。系统评估概率

$$P\left(ApplyAbs\left(ev_j(\tau+\Delta), abs(f_k, en)\right), gen(f_k)(ev_i(\tau))\right).$$

在阶段1-2，系统构建一个分类器，估计因果事件的词法特征的任何实体先于故事情节中事件的文本中目标事件的出现的概率。例如，在这个阶段，最佳分类器将根据基加利(卢旺达首都)这个实体对“卢旺达霍乱爆发”的概率进行估计。在第3阶段，算法迭代地估计给定任何抽象特征(使用分层图G^H提取)的目标事件发生的概率。例如，一个迭代可以是评估在我们的故事情节中"非洲首都"实体先于"卢旺达霍乱爆发"实体的次数。阶段3.1-3.2在给定抽象原因实体的情况下，评估到目标事件所需的转换。例如，我们不是寻找发生了属于“非洲首都”实体的事件，然后发生了与“卢旺达霍乱爆发”有关的事件的案例，而是寻找随后发生了“非洲霍乱爆发”类型事件的案例。然后，我们使用转换后的训练数据来训练和评估新的分类器。如果它的性能(通过对训练数据的交叉验证来衡量)优于抽象之前的分类器，我们就更新找到的最佳分类器。

3. 实验评价

我们现在描述了我们为测试方法而进行的实验，并展示了对训练阶段的新闻档案的测试部分进行的推断研究的结果。

3.1 实验装置

在本节中，我们概述了我们为实验获得的数据，实验方法以及我们比较的基线。

3.1.1 数据

我们抓取并解析了1986年至2007年期间《纽约时报》的新闻文章。我们说事件链属于一个域D，如果它包含一个域相关的词，表示为w_i(D)。例如，对于预测未来死亡的挑战，我们考虑单词“killed”、“dead”、“death”以及它们的相关术语。[3]对于预测未来疾病爆发的挑战，我们考虑所有提到“霍乱”、“疟疾”和“痢疾”的词。

在预测过程中，我们从学习阶段拿出1998-2007年(测试期)十年事件的测试集。如果(1)它的所有事件的日期都发生在测试周期日期，并且(2)链中的第一个时间顺序事件不包含一个领域术语，例如，第一个事件不包含提到死亡(否则预测可能是微不足道的)，我们说链是一个测试域链。形式上，设C = {e_1. .e_k}是一个测试链，因此∀i:w_i(D)6∈e_1。

[3]We consider all the similarity relations in Wordnet: Synonyms, pertainyms, meronyms/ holonyms, hypernyms/hyponyms, similar to, attribute of, and see also relations.

Procedure ABSTRACT$(target, cause, Chains, G^H, G, k)$
(1) **Foreach** $\{entity \in Entities(cause)\}$
(1.1) $PositiveExamples \leftarrow \{(ev_1, ev_2) | ev_1 <<_{c \in Chains} ev_2, entity \in ev_1, \forall e \in Entities(target) : e \in ev_2\}$
(1.2) $NegativeExamples \leftarrow \{(ev_1, ev_2) | ev_1 <<_{c \in Chains} ev_2, entity \in ev_1, \exists e \in Entities(target) : e \notin ev_2\}$
(2) $bestClassifier \leftarrow Build(PositiveExamples, NegativeExamples)$
(3) **Foreach** $\{entity \in Entities(cause), absEntity \in Abstractions(entity, G^H)\}$
(3.1) $absPaths \leftarrow FindPaths(absEntity, Entities(target), G, k)$
(3.2) $absTargets \leftarrow ApplyAbs(absEntity, absPaths, G)$
(3.2) **Foreach** $absTaret \in absTargets$
(3.2.1) $PositiveExamples \leftarrow \{(ev_1, ev_2) | ev_1 <<_{c \in Chains} ev_2, absEntity \in ev_1, \forall e \in Entities(absTarget) : e \in ev_2\}$
(3.2.2) $NegativeExamples \leftarrow \{(ev_1, ev_2) | ev_1 <<_{c \in Chains} ev_2, absEntity \in ev_1, \exists e \in Entities(absTarget) : e \notin ev_2\}$
(3.2.3) $absClassifier \leftarrow Build(PositiveExamples, NegativeExamples)$
(3.2.4) **If** $CV(bestClassifier, Chains) < CV(absClassifier, Chains)$
$bestClassifier \leftarrow Update(absClassifier)$
(4) **Return** bestClassifier

图3:通过抽象泛化特征的过程。Build将正反例作为输入，并估计我们目标事件的概率。FindPaths查找作为输入给定的图中两个节点之间大小为k的所有谓词路径。ApplyAbs在一个节点上应用谓词路径，返回通过有向路径的谓词连接到给定节点的节点。CV通过对训练数据进行分类器的交叉验证来计算精度。

3.1.2 实验方法

对于每个预测实验，我们首先从测试域链中选择一个目标事件e_{target}。根据实验类型的不同，过程有所不同:

1. 预测2006-2007年期间的一般事件。在这种类型的实验中，目标事件是2006-2007年期间发布的任何新闻标题，也就是说，我们为每个可能的标题构建一个分类器。

2. 预测具体三个领域的事件:死亡、疾病爆发和骚乱。在这种情况下，任何包含其中一个领域词的新闻故事都会被选中。此外，我们手动验证这些事件是否确实包含来自该领域的事件。如果存在多个目标事件，我们选择按时间顺序出现的第一个作为已识别的目标事件，即$e_{target}=argmin_j\{e_j|\exists i:w_i(D)\in e_j\}$。由于$e_{target}$是从测试域链j > 1中选择的，也就是说，它不是链中的第一个事件。也就是说，我们只考虑1998-2007年期间系统未观察到的事件链，并且在第一个事件链期间不包含暗示域内目标事件的单词(例如，单词death)。该链的第一个事件被作为系统的输入。

总而言之，一般事件预测代表了对2006-2007年所有事件的预测。系统给出一个2006-2007年的事件作为输入，我们衡量该事件预测的成功程度。对于特定领域的预测(死亡、疾病爆发和骚乱)，我们使用领域代表词或其同义词作为过滤器，手动检查事件是否发生。我们只考虑1998-2008年期间未观察到的事件链，并且在第一个事件链期间不包含暗示域内目标事件的单词(例如，单词death)。将该链的第一个事件作为输入。

我们通过评估在链中第一个事件发生日期之前发生的事件e_{target}发生的概率来从数据中进行训练。在测试过程中，算法呈现链e_1的第一个事件，并输出其对e_{target}的预测。在实验中，我们认为预测器表明目标事件将发生，如果

$$P(e_{target}|e_1) > P(\neg e_{target}|e_1),$$

也就是说，给定e_1事件发生的概率大于它不发生的概率。我们在所有相关链上重复执行这些实验，并对每个链进行评估:

$$\text{precision} = \frac{|\{\text{events reported}\} \cap \{\text{predicted events}\}|}{|\{\text{predicted events}\}|}$$

和

$$\text{召回(灵敏度)} = \frac{|\{\text{上报事件}\} \cap \{\text{预测事件}\}|}{|\{\text{事件报道}\}|}。$$

3.1.3 比较分析

我们不知道文献中有任何方法旨在解决对未来新闻事件概率的预测。因此，我们将生成的预测与两个基线进行比较:

1. 在给定训练集P(e)中对应文本的出现情况下，使用事件e发生的先验概率;

2. 使用对人们预测这些类型事件的能力的估计。

对于后者，我们实现了一种方法[11]，该方法提供了人们是否会同意第一个事件意味着后一个事件的真相的近似值，给定两个由新闻故事文本表示的事件。该基线评估文本中而不是时间中共现的概率。

		Predicted		
		Few	Tens	Hundreds
Real	Few	40%	6%	1%
	Tens	4%	32%	1%
	Hundreds	1%	6%	9%

表2:显示预测死亡人数与实际死亡人数的混淆矩阵。

General Predictions		Death		Disease Outbreak		Riots	
Med.	Avg.	Med.	Avg.	Med.	Avg.	Med.	Avg.
9	21	8	41	12	273	18	30

表4:基于推断的世界上结果和目标事件的概率的警报之间的中位数和平均时间(天)。

3.2 预测结果

我们进行了实验，评估了一般预测和每个不同领域的精度和召回率。我们将我们的模型(Full模型)与基于频率的模型(Frequency)和基于共发生的方法(Co-occurrence)进行了比较。结果如表1所示。我们观察到，在所有情况下，Full模型的表现都优于基线。

我们进行了额外的实验来评估数值预测，例如对死亡人数的预测。为此，我们在“[数字]死亡”或“[数字]死亡”形式的新闻报道中搜索特定模式。将匹配[number]的数字作为bin分类。我们只关注包含这些模式的链，并在这些链上评估我们的算法。在表2中，我们显示了死亡人数的混淆矩阵。每个单元格i、j的内容(i为行，j为列)表示实际属于bin i并被归类为属于bin j的数据的百分比。例如，4%导致数十人死亡的事件被错误地预测为仅与少数(少于10人)死亡相关的事件。我们在这些类型的分类中看到了很高的性能，并且在相邻的箱子中观察到大多数错误。

3.3 算法分析

我们现在描述了为测量程序的性能和分析的特定组件的贡献而进行的其他实验。

3.3.1 事实特征和泛化的增益

我们首先考虑添加不同世界知识来源对预测准确性的影响。结果如表3所示。我们考虑了仅基于词汇特征(单独的新闻)，同时基于词汇和事实特征(新闻+事实特征)，基于词汇特征和抽象(新闻+概括)，以及使用所有类别的特征和抽象过程(Full模型)的预测器。我们发现，无论是在抽象还是在添加事实特征时，添加知识都能提高预测的性能。我们发现，在同时使用这两种改进时，性能提升最大。

3.3.2 预测即将发生的事件的时间

表4显示了我们研究的三种预测类型中基于推理的警报与体现目标事件的报告发生之间的平均时间和中位数时间。我们只考虑死亡事件出现在新闻标题中并与手工制作的模板(“[number] died”或“[number] killed”形式的文本模式)相匹配的例子，以识别仅在测试链上的某些死亡事件。这一过程产生了951例死亡预测。在许多情况下，我们发现警报会比目标事件提前一周多。我们在图4中说明了这一现象，其中显示了对死亡人数的预测在警报和死亡事件发生之间的不同时间预测的故事情节内的未来时间。图5显示了事件发生前2天和15天的警报时间的更详细视图。

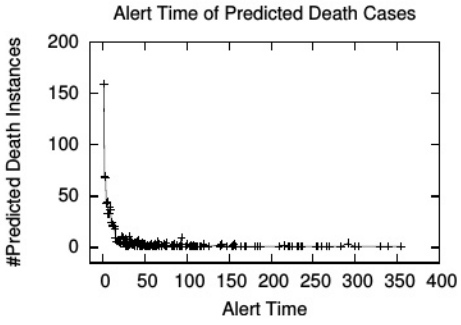

图4:预测任意数字的死亡次数作为警报时间(事件发生前几天)的函数。

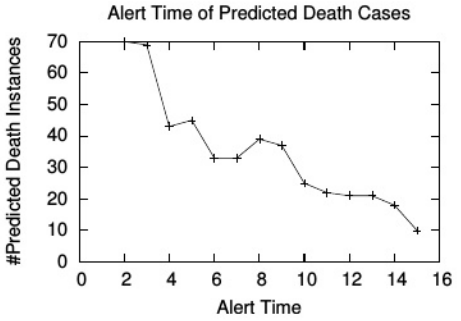

图5:预测任意数量的死亡次数作为警报时间(事件发生前几天)的函数。

3.4 事件链提取评估

为了评估提取的事件链的质量，我们在TDT4语料库[4]上进行了实验，只过滤了NYT文章。该语料库包含大约280,000篇文档，时间为04/01/2003-09/30/2003。人类标记故事情节的方法是由TDT挑战的组织者执行的。对于每个链，我们计算平均精度-我们提取的文章在链中确实是故事情节的一部分的百分比。我们还计算平均召回率，即系统实际检索到的文章链中的文章数量。我们将使用实体熵度量的事件链提取器与

[4] http://www.nist.gov/TDT

	General Predictions		Death		Disease Outbreak		Riots	
	Prec.	Recall	Prec.	Recall	Prec.	Recall	Prec.	Recall
Full model	24%	100%	83%	81%	61%	33%	91%	51%
Frequency	<1%	100%	59%	<1%	13%	3%	50%	1%
Co-occurrence	7%	100	46%	61%	40%	<1%	61%	14%

表1:几个领域预测的精度和召回率。

	General Predictions		Death		Disease		Riots	
	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.
News alone	19%	100%	80%	59%	44%	34%	88%	38%
News + factual features	19%	100%	81%	62%	52%	31%	87%	42%
News + generalization	21%	100%	81%	67%	53%	28%	88%	42%
Full model	24%	100%	83%	81%	61%	33%	91%	51%

表3:不同算法配置的精度和召回率。

在没有熵测度的情况下工作的提取器。结果总结在表5中。结果表明，虽然文本聚类的召回率非常高(提高10%)，但精度明显低于我们提出的方法(降低30%)。因此，我们更倾向于第二种方法，因为它在以更精确的方式训练预测模型方面提供了更大的灵活性，并影响了用于训练学习者的示例数量。

	Precision	Recall
Text Clustering	34%	**80%**
Text Clustering + Entity Entropy	**70%**	63%

表5:链提取过程的精度和召回率。

3.5 样本可能性和故事情节

我们描述的学习和推理方法可用于从观察序列中输出感兴趣的关键转换的概率。系统通过网络上的新闻和相关数据的更新来不断完善其学习。正如我们提到的，系统可以根据它正在监控的特定结果集的新闻报道提供实时警报。图6显示了代表性学习转移概率的统计示例。这些过渡概率和平均过渡时间突出了方法提供关于各种抽象级别的推论的能力。

现在我们将详细介绍几个额外的故事情节，以及推论和时间安排。考虑图1中图形化显示的示例。2007年1月26日，《纽约时报》发表了一篇关于非洲风暴和洪水的文章。四天后，霍乱疫情的消息被报道出来。针对这一连串的新闻，我们描述的方法产生了两个警报，一个是在观察2006年初安哥拉的干旱报告时，另一个是在报道风暴的新闻之后。该系统从其训练集中的许多类似事件中了解到，干旱后霍乱爆发的可能性更高，特别是因为干旱观测报告与后来与水有关的灾害报告的可能性增加有关，而后者又与水传播疾病报告的可能性增加有关。这种转变和可能性的例子包括该系统分析的孟加拉国的一系列干旱。1960-1991年期间，孟加拉国报告了19次重大干旱[19]。我们观察到，在描述这些干旱的故事情节中，在故事情节的后期，84%的病例报告了霍乱爆发。1973年的干旱导致了1974年的饥荒，1975年10月13日，《NYT》报道:“霍乱疫情袭击孟加拉国;可能比1974年创下记录的那次更严重……”。1983年3月13日，也就是1982年旱灾的一年后，“造成稻米减产约53000吨，而同一年，洪水破坏了36000吨……”，《NYT》发表了题为"孟加拉国霍乱死亡"的文章。几个月后，出现了一篇题为"据报霍乱在孟加拉国三次爆发中造成500人死亡"的文章。根据这些过去的故事情节，系统推断出2007年1月底至1月底爆发的霍乱。

预测方法了解到，并非所有的干旱都与此类疾病爆发可能性的跳跃有关。特定的前提条件会影响从干旱报告转变为霍乱爆发报告的可能性。该方法能够认识到，1989年3月纽约市发生的干旱与疾病爆发无关，该事件发表在《NYT》上，标题为:"因干旱而宣布紧急状态"。唯一的结果是纽约市宣布限水，并于当年5月16日结束。该系统估计，要使干旱极有可能引起霍乱，干旱必须发生在靠近水体的不发达国家人口稠密的地区(如安哥拉和孟加拉国的难民营)。

作为预测的另一个例子，我们着重讨论1991年孟加拉国霍乱流行的情况。据估计，这次霍乱爆发包括21万例霍乱病例，8000多人死亡[16]。在我们的实验中，我们发现运行中的预测系统会在霍乱爆发开始前4天发出警报，随后观察到主要的洪水。在图6中，我们以图形方式显示了检测到的故事线。该系统确定，在孟加拉国，高概率的大洪水报告之后，将出现重大疾病暴发的报告。该系统的推论得到了一项关于孟加拉国霍乱流行的大型研究的支持[16]，该研究基于对政府数据的分析，以及1985年至1991年间在孟加拉国400个农村地区独立收集的数据。分析表明，1987年和1988年的霍乱病例和死亡人数明显高于其他年份(300,000-1,000,000例，而其他年份为50,000例)。1987年和1988年，孟加拉国发生了严重的洪水。该研究得出结论，

Cause	Effect	Probability
Drought	Flood	18%
Flood	cholera	1%
Flood in Rwanda	cholera in Rwanda	67%
Flood in Lima	cholera in Lima	33%
Flood in Country with water coverage > 5%	cholera in Country	14%
Flood in Country with water coverage > 5%, population density > 100	cholera in Country	16%

表6:几个例子的概率转换。

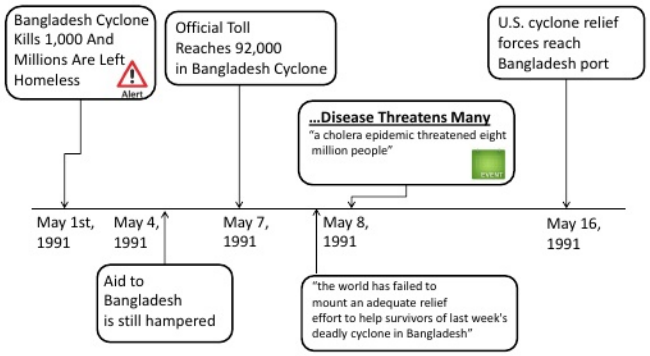

图6:孟加拉国风暴后霍乱警报的例子。三角形警报图标表示即将到来的霍乱爆发可能性显著上升的推断。

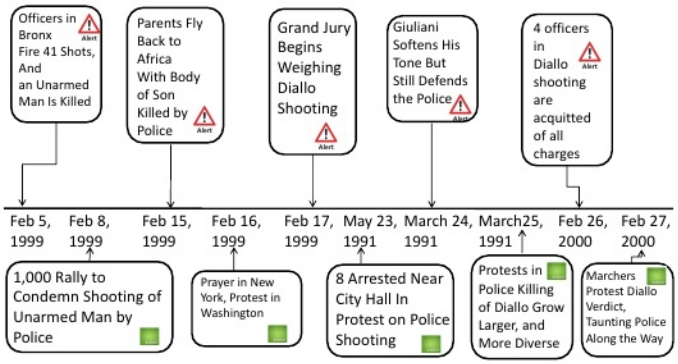

图7:手无寸铁的少数民族被枪杀后即将发生骚乱的可能性警报示例。三角形的警报图标代表了即将发生骚乱的可能性显著上升的推断。

获得医疗服务是许多非农村地区死亡率高的主要原因之一。在早期阶段就采取适当干预措施的地区，死亡率要低得多。

我们还研究了先前死亡和骚乱的例子。在图7中，我们给出了1999年Dialo案件的部分故事情节和推断的警报。其中一个故事情节在表7(上图)中有详细的描述。该系统以自动化的方式识别出，对于移民人口众多的地区(如俄亥俄州和纽约)，警察射杀手无寸铁的人可能会引发抗议。新闻中的其他事件，如以类似方式被杀害的人的葬礼的报道，对开枪的警察的审判开始的报道，对警察的支持，以及审判结束的报道，都与后来报道抗议的可能性增加有关。表7(下)给出了一个基于推断概率的故事情节示例。

4. 相关工作

先前的相关研究包括政治学中从编码事件数据预测即将到来的国际政治危机的努力，包括从新闻故事中提取的事件数据[23]。这一领域的研究包括应用hmm来识别似乎与国际危机发展有关的属性之间的相似性[22]。相关研究还探索了从Twitter等社交媒体衍生的信号中预测骚乱[13]和电影票销售[5,12,18]。其他调查利用新闻和图书语料库文本中的信息，定性地估计了多个方面

人类文化进化[25,25,17]。在搜索和检索方面的其他相关工作集中在将传统媒体[20]和博客[1]中输入搜索引擎的查询日志与未来事件相关联。Ginsberg等人[10]使用查询来预测H1N1流感爆发。其他研究试图预测网络内容如何变化。Kleinberg[14,15]开发了通用技术，用于总结文本内容的时间-门户动态和识别内容中的术语爆发。类似地，其他作品[4]在与查询相关的文档的发布日期上构建时间序列模型，以预测未来的突发事件。在其他相关工作中，Radinsky等人[21]从过去的新闻中以“x导致y”的形式提取了广义模板。将模板应用于当前的新闻标题，生成一个似是而非的未来新闻标题。

在这项工作中，我们采用概率方法，在不依赖模板的情况下执行更通用的预测。我们还结合了异构的在线资源，利用从网络上90多个来源中挖掘的世界知识，丰富和概括历史事件，以预测未来的新闻。

5. 结论

我们提出了从22年的新闻档案中挖掘事件链的方法，以提供一种方法，可以实时预测未来感兴趣的世界事件的可能性。该系统利用多个Web资源来概括它所学习和预测的事件。我们讨论了如何从大量数据中学习模式，监控大量信息源，并继续学习新的概率关联

Date	Title
Jan 16, 1992	Jury in Shooting by Officer Hears Conflicting Accounts
Feb 11, 1992	Closing Arguments Conflict on Killing by Teaneck Officer
Feb 12, 1992	**[Past Event] Officer Acquitted in Teaneck Killing**
Feb 13, 1992	Acquitted Officer Expresses Only Relief, Not Joy
Feb 16, 1992	**[Past Riot] 250 March in Rain to Protest Teaneck Verdict**

Feb 24, 2000	Diallo Jurors Begin Deliberating In Murder Trial of Four Officers
Feb 26, 2000	**[Riot Alert] 4 officers in Diallo shooting are acquitted of all charges**
Feb 26, 2000	Rage Boils Over, and Some Shout 'Murderers' at Police
Feb 26, 2000	Civil Rights Prosecution Is Considered
Feb 27, 2000	**[Riot Event] Marchers Protest Diallo Verdict...**
Feb 27, 2000	2 jurors defend Diallo acquittal

表7:上表:用于推断概率的历史故事线的部分样本。下表:带有警报的部分故事情节。

ciations。为了演示该方法，我们展示了几个评估的结果以及事件序列和主动警报的代表性示例。我们考虑了关于疾病爆发、骚乱和死亡的预测作为样本推论。我们认为，这些方法突出了构建实时警报服务的方向，这些服务可以预测全球感兴趣的事件的显著增加。除了在研究中容易发现或从专家那里获得的知识之外，还可以通过这种自动化分析发现结果的新关系和上下文敏感概率。采用这些方法的系统将能够快速、全面地访问新闻故事，包括那些看似无关紧要但可以为更大、更重要的故事的演变提供有价值证据的故事。我们希望这项工作将激发更多的研究，利用过去的经验和人类知识，为未来事件和重要干预提供有价值的预测。

6. 引用

[1] E. Adar, D. S. Weld, B. N. Bershad, and S. D. Gribble. Why we search: visualizing and predicting user behavior. In WWW, 2007.

[2] A. Ahmed, Q. Ho, J. Eisenstein, E. Xing, A. J. Smola, and C. H. Teo. Unified analysis of streaming news. In Proc. of WWW, 2011.

[3] J. Allan, editor. Topic detection and tracking: event-based information organization. Kluwer

Academic Publishers, Norwell, MA, USA, 2002.

[4] G. Amodeo, R. Blanco, and U. Brefeld. Hybrid models for future event prediction. In CIKM, 2011.

[5] S. Asur and B. A. Huberman. Predicting the future with social media, 2010.

[6] C. Bizer, T. Heath, and T. Berners-Lee. Linked data – the story so far. IJSWIS, 2009.

[7] J. Carbonell, Y. Yang, J. Lafferty, R. D. Brown, T. Pierce, and X. Liu. Cmu report on tdt-2:

segmentation, detection and tracking, 2000.

[8] C. Cieri, D. Graff, M. Libermann, N. Martey, and S. Strassel. Large, multilingual, broadcast news corpora for cooperative research in topic detection and tracking: The tdt-2 and tdt-3 corpus efforts. In LREC, 2000.

[9] J. R. Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In Proceedings of ACL, 2005.

[10] J. Ginsberg, M. Mohebbi, R. Patel, Brammer, M. L., Smolinski, and L. Brilliant. Detecting influenza

epidemics using search engine query data. Nature, 457:1012−1014, 2009.

[11] O. Glickman, I. Dagan, and M. Koppel. A probabilistic classification approach for lexical textual entailment. In Proc. of AAAI, 2005.

[12] M. Joshi, D. Das, K. Gimpel, and N. A. Smith. Movie reviews and revenues: An experiment in text

regression. In In Proc. of NAACL-HLT, 2010.

[13] Kalev. Culturomics 2. 0: Forecasting large-scale human behavior using global news media tone in time and space. First Monday, 15(9), 2011.

[14] J. Kleinberg. Bursty and hierarchical structure in streams. In KDD, 2002.

[15] J. Kleinberg. Temporal dynamics of on-line information systems. Data Stream Management:

Processing High-Speed Data Streams. Springer, 2006.

[16] J. Michel, Y. Shen, A. Aiden, A. Veres, M. Gray, Google Books Team, J. Pickett, D. Hoiberg,

D. Clancy, P. Norvig, J. Orwant, S. Pinker,

M. Nowak, and E. Aiden. Cholera epidemics in

bangladesh: 1985-1991. Journal of Diarrhoeal Diseases Research (JDDR), 10(2):79−86, 1992.

[17] J. Michel, Y. Shen, A. Aiden, A. Veres, M. Gray, Google Books Team, J. Pickett, D. Hoiberg,

D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. Nowak, and E. Aiden. Quantitative analysis of culture using millions of digitized books. Science, 331:176−182, 2011.

[18] G. Mishne. Predicting movie sales from blogger sentiment. In In AAAI Spring Symposium, 2006. [19] R. Nagarajan. Drought Assessment. Springer, 2009.

[20] K. Radinsky, S. Davidovich, and S. Markovitch. Predicting the news of tomorrow using patterns in web search queries. In WI, 2008.

[21] K. Radinsky, S. Davidovich, and S. Markovitch. Learning causality for news events prediction. In Proceedings of WWW, 2012.

[22] D. Richards, editor. Political Complexity: Nonlinear Models of Politics. Ann Arbor: University of Michigan Press, Norwell, MA, USA, 2000.

[23] R. J. Stoll and D. Subramanian. Hubs, authorities, and networks: Predicting conflict using events data, 2006.

[24] F. Suchanek, G. Kasneci, and G. Weikum. Yago: a core of semantic knowledge. In Proc. of WWW, 2007.

[25] C. Yeung and A. Jatowt. Studying how the past is remembered: Towards computational history through large scale text mining. In Proc. of CIKM, 2011.