一种基于用户兴趣的微博实体链接方法*

宋 俊[1,2,3]，李禹恒[1,2,3] 黄 宇[1,2] 陈 昊[1,2,4] 付 琨*[,1,2]

（1. 中科院空间信息处理与应用系统技术重点实验室，北京 100190；2. 中国科学院电子学研究所，北京 100190；3. 中国科学院大学，北京 100190；4. 北京空间信息中继传输研究中心）

摘 要：本文针对微博内容较短、歧义较大的问题，利用概率主题模型对用户的兴趣进行建模，提出了一种基于用户兴趣的微博实体链接方法。具体地，本文首先利用现有的主题模型从知识库的大量数据中训练实体与上下文词汇的语义关联，然后提出用户兴趣主题模型来建模用户对实体的兴趣以及微博的语义，并完成实体链接的任务。此外，本文在真实数据集上进行了大量实验和分析，取得了87.6%的实体链接准确率，实验结果表明，与现有方法相比，该方法通过用户兴趣的建模更好地刻画了微博的语义，因而也取得了更高的实体链接准确率。

关键词：自然语言理解；实体链接；实体消歧；概率主题模型；用户兴趣建模

中图分类号： TP181 **文献标识码**：A **文章编号**：

A user interest topic model for microblog entity linking

SNOG Jun[1,2,3], LI Yuheng[1,2,3], HUANG Yu[1,2], CHEN Hao[1,2,4], FU Kun[1,2]

(1.CAS Key Laboratory of Spatial Information Processing and Applied System Technology, Beijing 100190, China; 2. Institute of Electronics, Chinese Academy of Sciences, Beijing 100190, China; 3. University of Chinese Academy of Sciences, Beijing 100190, China; 4. Beijing space information relay transmission technology research center)

Abstract: The objective of this paper is to propose a novel entity linking method for microblog based on user's interest. The proposed method models user's interest to enrich the semantics of microblog. To achieve this, this paper firstly utilizes an existing topic model to train the relatedness between entities and context words from Wikipedia. Then it proposes a user interest topic model to capture user's interest on entities and represent microblog's semantics. And the linking entity is also obtained from this topic model. Moreover, an efficient Gibbs sampler is developed. Our experimental results on real-world dataset verify that the proposed method can better reflect the semantics of microblog and also outperforms the baseline method in terms of linking accuracy.

Key words: natural language process; entity linking; entity disambiguation; topic model; user interest model

0 引言

随着人类在互联网上产生数据的日益增多，数据的分析和管理面临巨大挑战，尤其是文本数据。文本数据的内容通常包含大量的命名实体（人名、地名、机构名等），这些命名实体对文本数据的语义分析与建模具有重要作用，但它们往往具有歧义性，比如，“他今天遇到偶像迈克尔乔丹”这句话，名称“迈克尔乔丹”可能指篮球运动员乔丹，也可能指大学教授乔丹，还可能是其他名叫乔丹的人。

实体链接（Entity Linking）[1]是对命名实体名称进行分析，发现名称在知识库中对应的无歧义对象，如果知识库不存在相应对象，则实体名称指向空（NIL）。实体链接对信息的发现和利用具有重要意义[2, 3]，可以对查询语句的实体进行消歧，从而改善搜索体验，还可以在问答系统对实体识别，更好地理解问题的意图。目前实体链接已经取得了较大的进展[4, 5]，这些方法大多分为两个步骤，首先从知识库中提取每个名称可能指代的实体，然后利用上下文相关性对实体进行排序。微博由于内容较短，上下文信息缺乏，还存在较多困难。一些

收稿日期:2015-02-15；**修改日期**：xxxx-xx-xx **基金项目**：863 国家高技术研究发展计划（2012AA011005）

作者简介：宋俊（1987-），男，湖北襄阳人，博士研究生，研究方向为文本挖掘（songjun210@mails.ucas.ac.cn）；李禹恒（1989-），男，吉林长春人，硕士研究生，研究方向为文本挖掘、自然语言理解；黄宇（1981-），男，辽宁丹东人，副研究员，研究方向为地理空间信息挖掘与可视化；陈昊（1982-），男，安徽铜陵人，工程师，研究方向为信号与信息处理；付琨（1974-），男（通信作者），湖北荆州人，研究员，研究方向为计算机视觉与遥感图像理解、地理空间信息挖掘与可视化（kunfuiecas@gmail.com）。

方法提出利用更多的相关微博或微博特征来进行链接[6, 7]，然而，这些方法大多基于词向量空间来进行语义表示，没有考虑词汇之间的关联，不能较好地刻画微博的语义。

概率主题模型是近年来提出的一种语义建模框架，广泛应用于文本数据分析，它假设文档是主题的分布，主题是词的分布[8]。一些方法提出利用概率主题模型来建模名称和实体的上下文语义，更好地刻画它们的相关性[9, 10]，概率主题模型在文本语义建模上具有较大的优势，但是，目前还没有合适的针对微博实体链接的概率主题模型。

本文首先利用条件独立主题模型[11]来训练实体与上下文词汇的语义关联，然后从微博的用户特性出发，提出了一种用户兴趣主题模型来联合建模用户的兴趣和微博的语义，用户的兴趣表示为主题的分布和实体的分布，并通过一个概率图模型来完成实体链接的任务。实验结果表明，该模型相比现有方法可以更好地刻画微博上下文与实体的相关性，并取得了更好的链接准确率。

1 基于用户兴趣的微博实体链接方法

本文提出的基于用户兴趣的微博实体链接方法主要包括三个步骤，首先从知识库提取名称与候选实体的映射表及相应的统计信息，然后利用知识库对候选实体的语义进行建模，最后分析名称与候选实体的语义，完成实体链接。

1.1 提取候选集合

提取名称与候选实体的映射表是对给定的每个名称，从知识库中提取出这个名称可能指代的候选实体集合。类似大多数研究工作[4, 12]，本文采用维基百科作为知识库数据的来源，在提取名称的候选实体集合过程中，主要考虑实体页面、重定向页面、歧义页面、页面超链接、实体属性等信息。

本文采用维基百科 2014 年 3 月 4 日版本作为实验知识库，通过对以上结构信息的提取和汇总，得到每个名称的候选实体集合。表 1 中展示了部分名称的候选实体映射表，对于每个名称，提取可能指代的候选实体以及相应的指代次数，次数反映了不同实体对于给定名称的流行强度。比如 Michael Jordan 指代篮球运动员的次数要远大于其他实体。

表 1 名称候选实体映射表

名称	候选实体	指代次数
Michael Jordan	Michael Jordan	954
	Michael I. Jordan	8
	Michael Jordan (mycologist)	5
	Michael Jordan (footballer)	4
	…	…
Java	Java	3622
	Java (programming language)	2472
	Java (software platform)	376
	Java (town)	15
	Java, New York	12
	…	…

通过表中名称的候选实体集合信息，可以得到以下概率信息：

$$P(e \mid m) = \frac{count(m,e)}{count(m,*)} \tag{1}$$

$$P(m \mid e) = \frac{count(m,e)}{count(*,e)} \tag{2}$$

其中，$count(m,e)$ 表示名称 m 指代实体 e 的次数，$count(m,*)$ 表示名称 m 的总次数，$count(*,e)$ 表示实体 e 被指代的总次数，$P(e \mid m)$ 反映了给定名称 m 的情况下，不同候选实体 e 的流行程度，$P(m \mid e)$ 反映了实体 e 产生名称 m 的概率，可用来衡量名称对候选实体的流行程度。

1.2 建模实体语义

每个实体都具有自己的上下文语义，同一名称在不同的语义环境中往往指代不同的实体。比如名称 Michael Jordan（迈克尔乔丹），在篮球领域很可能指代篮球运动员，而在教育领域则更可能指代大学教授。建模实体语义是利用知识库中对实体的描述信息建立实体的上下文语义，这样才能在链接过程中对实体语义进行比较和分析。

概率主题模型在文本语义建模上具有较大优势，同时，维基百科的实体描述信息，还包含了丰富的超链接信息，可以对实体的语义进行增强。因此，本文采用 Newman[11]提出的条件独立主题模型（Conditionally-Independent Latent Dirichlet Allocation, CI-LDA）来建模实体语义，该模型同时利用维基百科页面的文本内容和超链接进行建模，CI-LDA 模型假设每个主题同时具有一个词分布 ϕ 和实体分布 $\tilde{\phi}$，每篇文档具有一个主题分布 θ，文档中的词和实体根据 θ 从相应的主题中产生，图模型如图 1 所示。

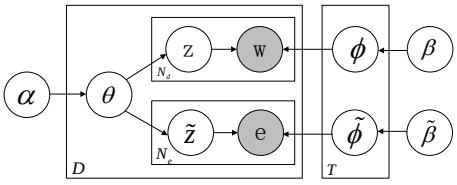

图 1 CI-LDA 图模型

CI-LDA 模型的产生式过程如下：

(1)对每个主题 $k \in K$：

(a)生成词在主题上的分布 $\phi \sim Dir(\beta)$

(b)生成实体在主题上的分布 $\tilde{\phi} \sim Dir(\tilde{\beta})$

(2)对每篇文档 $d \in D$：

(a)生成文档的主题分布 $\theta_d \sim Dir(\alpha)$

(b)对文档的每个词 $i \in N_d$：

i. 选择词的主题分配 $z_{d,i} \sim Mult(\theta_d)$

ii.选择词 $w_{d,i} \sim Mult(\phi_z[z = z_{d,i}])$

(c) 对文档的每个实体 $j \in N_e$：

i. 选择实体主题 $\tilde{z}_{d,j} \sim Mult(\theta_d)$

ii. 选择实体 $e_{d,j} \sim Mult(\phi_{\tilde{z}}[\tilde{z} = \tilde{z}_{d,j}])$

本文采用 Gibbs 采样算法对 CI-LDA 模型进行参数推断，其中，变量 θ, ϕ 和 $\tilde{\phi}$ 可以通过共轭特性得到，所以只需对隐变量 z 和 $\tilde{z}$ 进行采样，公式如下：

$$P(z_{d,i} = k \mid \cdots) = \frac{\alpha + n_{d,k\backslash i}}{K\alpha + n_{d,*\backslash i}} \frac{\beta + n_{k,w_{d,i}}}{V\beta + n_{k,*}} \quad (3)$$

$$P(\tilde{z}_{d,j} = k \mid \cdots) = \frac{\alpha + n_{d,k\backslash j}}{K\alpha + n_{d,*\backslash j}} \frac{\tilde{\beta} + \tilde{n}_{k,e_{d,j}}}{\tilde{V}\tilde{\beta} + \tilde{n}_{k,*}} \quad (4)$$

其中，公式中 $n_{d,k\backslash i}$ 表示文档 d 中除了词 $w_{d,i}$ 之外选择主题 k 的词和实体的个数，$n_{d,*\backslash i}$ 表示文档 d 中除了词 $w_{d,i}$ 之外的词和实体的个数，$n_{d,k\backslash j}$ 和 $n_{d,*\backslash j}$ 与此类似，$n_{k,v}$ 和 $\tilde{n}_{k,\tilde{v}}$ 分别表示主题 k 中词 v 和实体 $\tilde{v}$ 的使用次数，$n_{k,*}$ 和 $\tilde{n}_{k,*}$ 分别表示主题 k 中词和实体的总使用次数。对变量 z 和 $\tilde{z}$ 迭代采样直至收敛，最后根据统计量得到主题的词分布和实体分布：

$$\phi_{k,v} = \frac{\beta + n_{k,v}}{V\beta + n_{k,*}}, \quad \tilde{\phi}_{k,\tilde{v}} = \frac{\tilde{\beta} + \tilde{n}_{k,\tilde{v}}}{\tilde{V}\tilde{\beta} + \tilde{n}_{k,*}} \quad (5)$$

CI-LDA 模型可以对知识库中的实体描述信息进行语义表示，利用主题把语义相近的实体和词联系在一起，从而在实体链接过程中，可以通过文本的词汇分析文本语义，然后根据语义对实体进行消歧和链接。

1.3 微博实体链接

微博数据由于内容较短，语义信息缺乏，传统的主题模型不能很好地建模微博语义。Zhao 于 2011 年提出 twitter LDA[13]，利用主题的分布来建模用户的兴趣，达到增强微博语义的目的。基于此，本文提出了用户兴趣主题模型（User Interest Topic Model, UITM），利用主题的分布和实体的分布来建模用户的兴趣，从不同尺度层面刻画用户兴趣，并把实体链接任务融入图模型，UITM 图模型如图 2 所示。

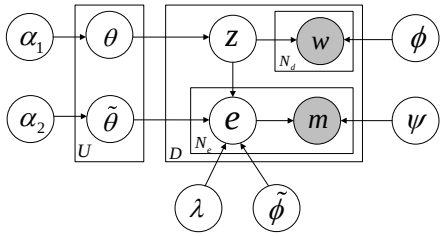

图 2 UITM 图模型

UITM 模型的产生式过程如下：

(1)对每个用户 $u \in U$：

(a)生成用户兴趣主题分布 $\theta_u \sim Dir(\alpha_1)$

(b)生成用户兴趣实体分布 $\tilde{\theta}_u \sim Dir(\alpha_2)$

(2)对用户 u 的每篇微博 $d \in D$：

(a)生成微博的主题分布 $z_d \sim Mult(\theta_u)$

(b)对每个词 $i \in N_d$，选择 $w_{d,i} \sim Mult(\phi_{z_d})$

(c)对微博的每个名称 $j \in N_e$：

i. 选择实体 $e_{d,j} \sim Mult(\lambda\tilde{\phi}_{z_d} + \tilde{\theta}_u)$

ii.选择名称 $m_{d,j} \sim Mult(\psi_e)$

在 UITM 模型中，ϕ 和 $\tilde{\phi}$ 分别表示 CI-LDA 模型中训练出来的主题的词分布和实体分布，ψ 表示从知识库中得到的名称与实体的先验知识，具体地，$\psi_{e,m}$ 表示在知识库中实体 e 产生名称 m 的概率，由公式（2）计算得到。用户兴趣体现在用户的主题分布 θ 和实体分布 $\tilde{\theta}$，前者从词的角度进行建模，后者从实体的角度进行建模，最终实体的产生同时依赖于两者，λ 表示对训练主题的实体分布 $\tilde{\phi}$ 的置信程度。

本文采用 Gibbs 采样算法对 UITM 模型进行参数推断，其中 θ 和 $\tilde{\theta}$ 可以根据共轭特性得到，所以需要采样的隐变量包括 z 和 e，采样公式如下：

$$P(z_d = k \mid \cdots) = \frac{\alpha_1 + n_{u,k\backslash d}}{K\alpha_1 + n_{u,*\backslash d}} \prod_{i=1}^{N_d} \phi_{k,w_{d,i}} \prod_{j=1}^{N_e} \tilde{\phi}_{k,e_{d,j}} \quad (6)$$

$$P(e_{d,j}=e\,|\cdots)=\frac{\lambda\tilde{\phi}_{k,e}+\alpha_2+n_{u,e\backslash j}}{\lambda+K\alpha_2+n_{u,*\backslash j}}\psi_{e,m_{d,j}} \quad (7)$$

其中，$n_{u,k\backslash d}$ 表示用户 u 除微博 d 之外选择主题 k 的次数，$n_{u,e\backslash j}$ 表示用户 u 除名称 $m_{d,j}$ 之外选择实体 e 的次数。对 z 和 e 进行迭代采样直至收敛，最终得到每篇微博中每个名称对应的实体 $e_{d,j}$，即实现了实体链接。

UITM 模型本身并没有很好地考虑 NIL 问题，本文利用公式（7）对每个实体的链接进行评分，然后通过交叉验证的方法学习得到一个阈值 τ 来判定给定的链接是否为 NIL。

实体之间存在着关联关系，这种关系对用户兴趣的建模具有重要作用。比如已经知道了用户对实体 Michael Jordan（篮球运动员）感兴趣，那么很容易推断出用户对实体 National Basketball Association（美国职业篮球协会）也会感兴趣。但 UITM 模型中用户兴趣的实体分布表述为一个多项式分布，实体之间相互独立，忽略了实体之间的关联关系。因此，本文提出利用实体的关联关系对用户的兴趣进行传播，可以在 UITM 模型的基础上方便地扩展实现。具体地，将公式（7）中的 $n_{u,e}$ 扩展为表示用户 u 对实体 e 的兴趣强度，计算公式如下：

$$n_{u,e}=\sum_{e'}c_{u,e'}r_{e',e} \quad (8)$$

其中 $c_{u,e'}$ 表示用户 u 选择实体 e' 的次数，$r_{e',e}$ 表示实体 e' 和实体 e 之间的关联关系，值越大表示实体之间关联越大。Milne 提出了一种度量维基百科页面之间关联度的方法（Wikipedia Link-based Measure, WLM）[14]，本文利用 WLM 来度量实体之间的关联关系，计算公式如下：

$$r(e_1,e_2)=1-\frac{\log(\max(|E_1|,|E_2|))-\log(|E_1\cap E_2|)}{\log(|WP|)-\log(\min(|E_1|,|E_2|))} \quad (9)$$

其中，E_1 和 E_2 分别表示指向实体页面 e_1 和 e_2 的维基页面集合，WP 表示维基的全部页面集合。利用以上关联关系，进行用户兴趣传播，更好地刻画用户兴趣。

2 实验结果及分析

本节在真实数据集上对 UITM 模型的有效性进行检验。首先介绍实验的微博数据集，然后利用知识库对实体语义进行建模和分析，最后定量地对微博实体链接的准确率进行比较。

2.1 实验数据

本文采用 Shen[7]提供的 tweet 数据集进行实验，该数据集包含不同用户的数据，且标注了命名实体和链接实体真值，相关信息如表 2 所示。

表 2 tweet 数据集信息

用户数量	Tweet 数量	名称数量	可链接的名称数量	不可链接的名称数量
20	3818	2677	2239	438

利用提取的名称候选集合可以得到每个名称的候选集合，该数据集涉及两千多名称，对应的候选实体总共三万多个，因此，方便起见，本文只选择部分与这些实体相关的维基百科页面作为实体语义的训练数据，最终得到的训练数据如表 3 所示。

表 3 实体语义训练数据

维基页面数量	总的词数量	总的实体数量	词典大小	实体词典大小
498646	333058778	16740487	142323	150617

2.2 实体语义建模

CI-LDA 模型训练实体语义实验中，文档的主题先验 设置为 0.1，主题的词和实体分布的先验分别设置为 0.01 和 0.001，主题个数设置 100，训练迭代 2000 轮，得到的部分主题如表 4 所示。

表 4 实体语义训练得到的部分主题

主题	Top10 词	Top10 实体
健康	health, medical, disease, medicine, hospital, treatment, research, care, patients, drug	Physician, Cancer, World Health Organization, Food and Drug Administration, Medicine, DNA, Protein, HIV, Doctor of Medicine, Tuberculosis
足球	league, season, club, team, against, cup, football, played, match,two	Association football, Premier League, FA Cup, Rugby union, UEFA Champions League, Arsenal F.C., Manchester United F.C., England national football team, The Football League, Chelsea F.C.
音乐	song, album, music, released, single, number, video, chart, songs, top	Billboard Hot 100, Billboard (magazine), UK Singles Chart, Billboard 200, MTV, Rolling Stone, Music video, Pop music, YouTube, Recording Industry Association of America
IT	system, car, software, engine, data, computer, available, model, windows, internet	Microsoft, Microsoft Windows, Internet, Linux, Google, Apple Inc., Android (operating system), IBM, Iphone, Facebook

表 4 中展示了实体语义训练得到的部分主题，

每个主题利用高频词和高频实体来表示。可以看出，每个主题的高频词和实体具有较强的一致性，可以用来进行主题传递并实现实体链接。比如一篇微博中频繁出现 league（联盟），club（俱乐部），cup（奖杯）等词，那么这篇微博很可能是关于“足球”主题的，利用一致性可以预测这篇微博很可能出现实体 Association football（足球联赛），FA Cup（足总杯），Arsenal F.C.（阿森纳足球俱乐部），Manchester United F.C.（曼联足球俱乐部）等，再根据名称对应的候选集合很容易得出链接的实体。

因此，CI-LDA 模型通过主题把实体和词汇联系起来，为实体链接提供了语义基础。

2.3 微博实体链接

本文采用基于流行度的实体链接方法[1]作为对比方法，即直接选择流行度最大的候选实体作为链接实体，同时为了分析用户兴趣建模的不同对微博实体链接准确率的影响，本文比较了忽略用户兴趣传播的 UITM 模型（记为 UITM_N）和考虑用户兴趣传播的 UITM 模型（记为 UITM_T）。实验中，用户主题分布超参数 设置为 0.01，用户实体分布超参数 设置为 1e-8，训练主题的置信度超参数 设置为 1000。在模型迭代采样到 1000 轮之后，继续迭代采样 1000 轮，其中每隔 10 轮记录一次采样值，最后选择出现次数最多的实体作为最终的链接实体。实验得到的准确率对比如表 5 所示。

表 5 微博实体链接准确率对比

	可链接		不可链接		全部	
Popular	1783	0.796	311	0.710	2094	0.782
UITM_N	1997	0.892	312	0.712	2309	0.863
UITM_T	2033	0.908	313	0.715	2346	0.876

从表 5 中可以看出，相比基于流行度的方法，两个 UITM 模型都取得了较好的结果，充分说明了建模用户兴趣对微博实体链接的重要性。同时，还可以看出，UITM_T 模型相比 UITM_N 模型在准确率上也有所提高，证明了考虑用户兴趣传播可以促进实体的链接。然而，对于 NIL 问题，UITM 模型并没有取得明显提高。NIL 问题可以看成是分类问题，UITM 等图模型并不适合这样的问题。

因此，UITM 模型较好地利用了用户兴趣来提升微博数据的语义，在实体链接准确率，尤其是对可链接名称的实体链接，有了较大提高。

3 结束语

本文提出了一种基于用户兴趣的微博实体链接方法，解决微博内容较短、歧义较大的问题。该方法首先利用条件独立主题模型训练实体与上下文词汇的语义关联，然后提出了用户兴趣主题模型 UITM，并在真实数据集上进行了实验和分析，取得了 87.6%的准确率，实验结果表明 UITM 模型通过对用户兴趣的建模丰富了微博的语义，得到了更高的实体链接准确率。然而，由于现有数据集的限制和数据集标注的困难，本文并没有在更大的数据集上进行测试，因此，在未来的研究工作中将进一步对该方法进行测试和改进，使该方法更加实用化。

参考文献

[1] SHEN, W. and J. HAN, Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions. IEEE Transactions on Knowledge and Data Engineering, 2014. 27(2): p. 443-460.

[2] 邢富坤，基于维基百科的领域实体发现研究．计算机应用研究，2015．32(2)．

[3] 姚宇峰，一种新的重名消解算法在保险领域中的应用研究．计算机应用研究，2012．29(3)．

[4] SHEN, W., et al. LINDEN: linking named entities with knowledge base via semantic knowledge. in Proceedings of the 21st international conference on World Wide Web. 2012. ACM.

[5] GUO, Y., et al., Improving candidate generation for entity linking, in Natural Language Processing and Information Systems. 2013, Springer. p. 225-236.

[6] Guo, Y., et al. Microblog entity linking by leveraging extra posts. in Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. 2013.

[7] SHEN, W., et al. Linking named entities in tweets with knowledge base via user interest modeling. in Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. 2013. ACM.

[8] BLEI, D.M., A.Y. NG, and M.I. JORDAN, Latent dirichlet allocation. The Journal of machine Learning research, 2003. 3(Jan): p. 993-1022.

[9] HAN, X. and L. SUN. An entity-topic model for entity linking. in Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. 2012. Association for Computational Linguistics.

[10] LI, Y., et al. Mining evidences for named entity disambiguation. in Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. 2013. ACM.

[11] NEWMAN, D., C. CHEMUDUGUNTA, and P. SMYTH, Statistical entity-topic models, in Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining2006, ACM: Philadelphia, PA, USA. p. 680-686.

[12] 张海粟，马大明，and 邓智龙，基于维基百科的语义知识库及其构建方法研究．计算机应用研究，2011．28(8)．

[13] ZHAO, W.X., et al., Comparing twitter and traditional media using topic models, in Advances in Information Retrieval. 2011, Springer. p. 338-349.

[14] MILNE, D. and I.H. WITTEN. An Effective, Low-Cost Measure of Semantic Relatedness Obtained from Wikipedia Links. in In Proceedings of AAAI 2008. 2008.