基于新闻文章差异分析的多元化移动新闻App

**摘要** 首先，通过对前人工作的拓展，在以实体为导向的新闻分析的基础上，从观点差异、细节差异、因子覆盖度差异三个角度提出了新闻文章检索与排序的三种措施。本文利用这些措施，在移动设备上开发新闻app，帮助用户获取多样化的报道，提高新闻的理解。该系统的一个显著的特性是一个上下文感知的重新排序方法，它通过考虑访问历史来增强呈现给用户的新闻报道的多样性。实验结果证明了该方法的有效性。

关键词:新闻app·多样性·差异性分析·上下文感知重排·众源实验

1介绍

从某种意义上说，由于编辑和赞助者的意图，新闻从来就不是没有偏见的。为了帮助用户理解新闻事件，考虑新闻文章的多样性是很重要的[1,2,5,8]。随着智能手机和平板电脑的普及，新闻应用，即智能移动设备上的新闻阅读应用，得到了广泛的应用。然而，新闻应用程序通常每个主题只提供一篇文章。此外，由于环境和设备的限制，移动搜索比一般的web搜索更加困难。因此，用户可能会失去从多视角获取信息避免偏见印象的机会。

在本文中，为了提高用户在智能移动设备上阅读新闻的体验，提出了一种新颖的新闻app，其功能是帮助用户在新闻事件中寻找多样化的信息。

首先，本文提出了三个方法来搜索和排名新闻文章，通过扩展之前的工作[1]。在[1]中，基于用户调查，本文提出了四种排序方法(相关性、多样性、意见差异、细节性差异)。[1]中描述的实验结果表明，虽然文章之间的差异很重要，但它们首先应该是相关的。基于这一观察，本文对这些排名方法进行了改进。改进措施总结如下。

- DC(因子覆盖率的差异)是指在报道同一事件的两篇新闻文章中所描述的不同内容的程度。

- DO(意见分歧)是指两篇报道同一事件的新闻文章主观描述的差异程度。

- DD(细节差异)是两篇报道同一事件的文章的细节差异程度。

利用这些措施，开发了一个新闻系统，有效地为同一事件提供不同的报道，支持用户对智能移动设备上的新闻理解。本文的主要贡献如下:

-提出了三个以实体为导向的新闻排名方法，通过扩展之前的工作(第3部分)，通过关注新闻报道中的差异来帮助搜索和排名新闻文章。

-提出一个寻求多样性的新闻应用(第4部分)。作为该系统的显著特征之一，本文提出了一种上下文感知的重新排序方法，以增强向用户提供的新闻报道的多样性。(Sect.4.3)。

-进行众包实验，验证排名措施的有效性(5.1节)。仿真实验验证了该方法的有效性(第5.2节)。

2相关的工作

为了帮助用户更好地理解新闻文章，提出了可视化和突出新闻文章之间差异的新闻浏览系统。Ogawa et al.[3]研究了新闻文章之间的差异分析，他们将注意力集中在命名实体上，并提出了一种保持机制。它们提取被提及的涉众，并呈现基于描述极性构建的图。Ogawa等人的目标新闻是文本新闻，Xu等人研究的是多媒体新闻[4]上的利益相关者挖掘。

NewsCube[2]表示新闻事件的各个方面，并使用方面查看器表示这些方面，以促进对新闻的理解。TVBanc[5]基于主题结构的概念对新闻文章进行比较。TVBanc从各大媒体搜集相关新闻，从中提炼出成对的话题和观点，揭示某一新闻事件中新闻报道的多样性和倾向性。

相比之下，本文研究的是通过估计文章之间的差异来搜索和排序相关的新闻文章。本文的方法侧重于如何向用户提供不同于用户已经阅读过的报告。此外，该系统是第一次尝试在智能移动设备上有效提供多样化的新闻报道。

3新闻稿件的实体化排序测度

在之前的[1]工作中，实验结果显示，虽然文章之间的差异对于获取不同的信息是很重要的，但是一篇报道完全不同事件的新闻文章对于帮助用户理解新闻是没有用的。基于这一观察，本文对定义进行了细化，提出了三种面向实体的排序方法:DC(因子覆盖率差异)、DO(意见差异)和dd(细节差异)。

在这一节中，首先简要介绍了从新闻文章中提取实体和实体相关描述的方法(具体参见[1]和[3])。然后介绍了面向实体的排序方法。

3.1实体和实体相关描述的提取

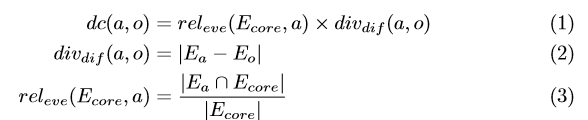
我们使用语言工具StanfordCoreNLP1提取命名实体，并应用Ogawaetal.[3]提出的一个方法生成一个树型结构(Ogawaetal称之为关系结构)。将对命名实体的描述看作关系结构的子树集，每个子树的根都是动词，其后代包含目标命名实体。

在面向实体的新闻分析中，另一个重要的概念是核心实体[1]。直观地说，事件中的核心实体是报告该事件的文章中经常提到的命名实体。

3.2排名措施

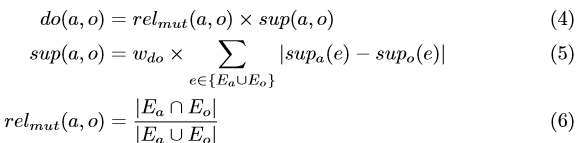
如前所述，面向实体的排名度量是基于比较新闻文章中的命名实体及其描述来定义的。

DC(因子覆盖率的差异):DC是指在报道同一事件的两篇新闻文章中描述的不同内容的程度。本文从两个方面估计DC，(1)这些文章中提到的不同因素有多少，(2)这些文章是否与同一事件相关。对于(1)，可以简单地比较文章中提到的实体。两篇文章描述的实体越多，两篇文章在因子覆盖率方面的差异就越大。对于(2)，将每篇文章中提到的实体与给定事件的核心实体集进行比较。让Ecore成为某个新闻事件的核心实体集。第a、o条之间的DC (a、o)计算如下:

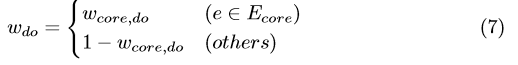


其中，Ea和Eo分别是a和o提到的命名实体集。

DO(意见分歧):表示两篇文章意见的不同程度。比较文章中命名实体的描述极性(正极性、负极性和中性极性)。如果两篇新闻报道的是同一个实体，而它们的极性不同，则认为这些文章是不同的意见。DO (a,o)计算如下:



其中，supa(e)和supo(e)分别为a和o中命名实体e的极性。wdo为核心实体的权重，计算方法如下:



DD(细节差异):DD表示两篇新闻文章提供的细节程度不同。比较了文章中与命名实体相关的描述，以估计DD在我们之前的工作中细节的差异。在提取命名实体相关描述之后，将LDA (potential Dirichlet Allocation)[6]应用于从命名实体描述中检测到的主题。然后，比较两篇新闻文章中每个命名实体的主题报道和主题词。DD包括权重wdd，计算方法与公式(7)相同。

4 .多元化的手机新闻App

利用提出的排名措施，开发了一个系统，帮助用户在智能移动设备上获取各种新闻事件的报道。

当用户在移动或等待过程中有少量时间时，用户可以使用移动设备查看新闻。然而，由于移动设备的时效性和特点，即使用户可能对某一新闻事件感兴趣，也很难找到关于该事件的各种报道。此外，典型的新闻应用程序每条新闻事件只显示一篇文章。这也是本文专注于智能移动设备新闻app的原因。本系统由新闻服务器和新闻客户端组成，主要功能见4.1 - 4.3节。

4.1新闻服务器:收集和分析新闻文章

本文的多元化新闻系统以英语新闻文章为目标。通过谷歌news Realtime Coverage2(谷歌news的功能之一，展示与头条新闻相关的文章)来收集头条新闻及其相关文章。新闻服务器在收集新闻之后，使用第3节中介绍的方法，对每个事件中的新闻文章进行分析，并将其排列为top news articles。新闻文章之间的差异分别用DC、DO和dd来量化。每篇头条新闻及其相关文章，每一项措施的最高排名，将被交付给新闻客户，然后呈现。当用户历史记录增加时，服务器将进行上下文感知的重新排序，如第4.3节所示。

4.2新闻客户端:新闻稿件展示

多元化新闻系统的新闻客户端有两种浏览模式。一个是今天的头条新闻。另一个是article details视图，它显示一篇新闻文章的内容，并链接到它的三篇排名最高的相关文章。图1展示了这两个视图的运行示例。

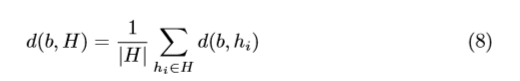


Top news视图显示从每个类别的谷歌新闻中收集的顶级新闻文章。用户可以单击一篇文章，在文章详细信息视图中查看其详细信息。

文章详细信息视图显示新闻文章的详细信息。有3个链接按钮分别命名为“Opposite”、“Wide”和“Deep”，分别对应DO、DC和DD中排名第一的文章。通过单击这些按钮，用户可以访问与当前事件相同的不同报告。

4.3环境敏感的评估

排名最靠前的文章可能是用户已经阅读过的文章，并且它们彼此相似，因为排名是通过与当前正在阅读的文章进行比较来决定的。为了从不同的文章中呈现更多不同的信息，本文提出了一种上下文感知的重排序方法。也就是说，比较的目标就是分级文章，不仅包括当前文章，还包括以前访问过的文章。假设是一组相关的新闻文章某一新闻事件和H(H⊂A)用户读过的文章。系统更新文章排名得分d(b、H)其中b∈{A−H }以H为跟进。



其中，d()表示dc()、do()和dd()。

5实验

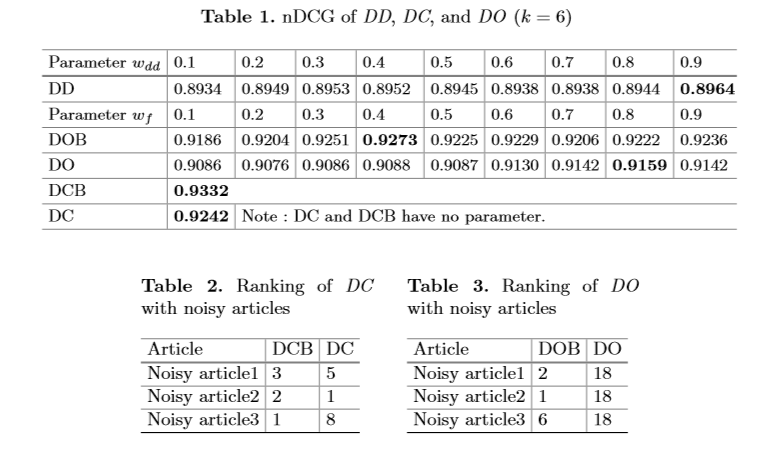
本文做了两个实验来评估提出的方法。一是排名标准。另一种是上下文感知的重新排序方法。

5.1众包排名实验

为了评估本文在不同人群中细化的排名措施，于是在CrowdFlower3提供的平台上进行了众源实验。在实验中，本文收集了2014年6月17日至24日谷歌news US edition (Top stories and the Realtime Coverage)的新闻文章。新闻事件数量为20篇，平均每篇新闻文章随机抽取14篇。在每个主题中，选择一篇文章作为当前文章，其他文章作为相关文章。在实验中，让工作人员比较这两篇文章和其中一篇相关文章。因此，平均每个新闻事件有13对比较。在每对比较中，分别从DC、DO、dd三个角度，按照五个等级评价体系对相关文章进行评分。

将众包工人的平均分数与提出的排名方法进行了比较。nDCG(归一化贴现累积收益)[7]是评价指标。改变每个参数wdo(公式7)和wdd (Sect.3.2)，计算20个主题(事件)排名前k (k = 6)的nDCG平均值。每个测度的nDCG结果如表1所示。

表1中，DCB和DOB为比较法，分别计算DC和DO，不考虑目标文章之间的相关性。这些分数的计算方法为DCB = divdif(a,o)和dob = sup(a,o)。

DD的nDCG结果中，最高评价值为0.8964,wdd = 0.9。这些值是可以接受的高。相反，很难说哪种方法更适合计算直流电，哪种方法更适合计算直流电。其中一个相当重要的原因是实验中使用的物品之间有很强的相关性。手动将一些有噪声的文章插入到数据集中的新闻事件中，并测试这些有噪声文章的等级。噪声文章的排名结果如表2和表3所示。从表中可以看出，DCB和DOB对噪声文章的排名较高，而DC对噪声文章的排名明显下降。因此，虽然该方法不能优于之前的方法，但它工作得很好，并且独立于相关新闻文章的收集准确性。

5.2情境感知重新排序方法实验

通过模拟新闻阅读序列，对情境感知的重新排序方法进行了实验评价。然后，随机选择了15个新闻事件用于众包实验。基于仿真的实验是对每个新闻事件重复以下步骤进行的。

1. 让a和ab成为当前的新闻文章。设A为某一新闻事件的相关文章集。让访问的文章H = a.

2。分别从DC, DO, DD三个方面对A中的文章进行排序，将这些排序分别存储在Rc, Ro, andd中。

3.执行上下文感知的重新排序方法，分别从三个方面对A中的文章进行排序。

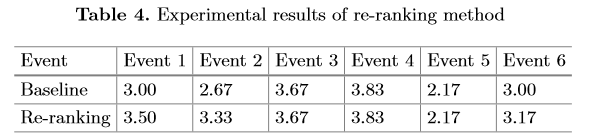
4. 随机选择一个方面x，但每个方面的选择不能超过两次。从选择的方面x中选择排名第一的文章t，从Rx中选择排名第一的文章tb。

5.一名评价者阅读文章t和tb，然后分别评估t和tb在5个等级中比较H有多少差异信息。

6. a ← t. ab ← tb. A = A−t, H = H∪t. Rc = Rc−tb,Ro = Ro−tb,Rd = Rd−tb.

7. 如果A不为空，则转到步骤3;其他的停止。

其中，a和t用于上下文感知的重新排序方法，ab和tb用于基线方法。

在用户调查[1]中，从相关性、观点、附加信息量三个角度，对新文章与之前阅读过的文章进行了5个等级的评价，以评估新文章与之前阅读过的文章相比，可以获得多少不同的信息。评价者从这三个角度对文章进行综合评分。基线有6个新闻事件的结果，重新排序的方法是不同的。计算了这六个事件的平均用户得分。结果如表4所示。 

在6个事件中，重新排序方法的用户得分大于或等于基线。这一结果揭示了重新排序方法的可用性，以增强用户阅读更多不同的信息。

6结论

在本文中，提出了三种面向实体的排名方法来支持用户获取同一新闻事件的不同报道。作为这些排名措施的应用之一，开发了一款智能移动设备上的多元化新闻应用。本文还提出了一种上下文感知的重新排序方法，以提供更丰富的信息。同时进行了众包实验来验证实体导向型排名措施。基于仿真的分析验证了上下文感知的重新排序方法。作为未来重要的工作之一，需要对新闻app进行用户研究和重新排名的方法。此外，计划利用本文的系统进行实验，研究用户媒介素养的变化。

确认。这项工作得到了KAKENHI(不是)的部分支持。25700033)和SCAT研究资助。

参考文献

1. Kiritoshi, K., Ma, Q.: Named entity oriented related news ranking. In: Decker, H., Lhotsk´a, L., Link, S., Spies, M., Wagner, R.R. (eds.) DEXA 2014, Part II. LNCS, vol. 8645, pp. 82–96. Springer, Heidelberg (2014)

2. Park, S., Kang, S., Song, S.C.J.: Newscube: delivering multiple aspects of news to mitigate media bias. In: CHI 2009, pp. 443–452 (2009)

3. Ogawa, T., Ma, Q., Yoshikawa, M.: News bias analysis based on stakeholder mining. IEICE Trans. 94–D, 578–586 (2011)

4. Xu, L., Ma, Q., Yoshikawa, M.: A cross-media method of stakeholder extraction for news contents analysis. In: Chen, L., Tang, C., Yang, J., Gao, Y. (eds.) WAIM 2010. LNCS, vol. 6184, pp. 232–237. Springer, Heidelberg (2010)

5. Ma, Q., Yoshikawa, M.: Topic and viewpoint extraction for diversity and bias analysis of news contents. In: Li, Q., Feng, L., Pei, J., Wang, S.X., Zhou, X., Zhu, Q.-M. (eds.) APWeb/WAIM 2009. LNCS, vol. 5446, pp. 150–161. Springer, Heidelberg (2009)

6. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993 (2003)

7. Jarvelin, K., Kekalainen, J.: Evaluation, cumulated gain-based, of IR techniques. ACM Trans. Inf. Syst. 20, 422–446 (2002)

8. Ma, Q., Nadamoto, A., Tanaka, K.: Complementary information retrieval for crossmedia news content. Inf. Syst. 31(7), 659–678 (2006)