文章编号: 1000-5641(2018)01-0091-12

基于前向分步算法的文档实体排序

王燕华

(华东师范大学 数据科学与工程学院, 上海 200062)

摘要: 文档中的关键实体可以抽象概括文本所描述的事件(或话题)的主体, 推动面向实体的检索和问答系统等方面的研究. 然而, 文档中的实体是无序的, 对文本中的实体进行排序显得尤为重要. 提取文本实体特征并借助维基百科和词汇分布表示引入外部特征, 提出了一种基于前向分步算法(Forward Stagewise Algorithm, FSAM)的排序模型 LA-FSAM(FSAM based on AUC Metric and Logistic Function). 该模型利用曲线下面积(Area Under the Curve, AUC)准则构造损失函数, 逻辑斯谛函数整合实体特征, 最后使用随机梯度下降法求解模型参数. 通过 LA-FSAM 与基线方法的实验对比证明了所提方法的有效性.

关键词: 实体排序; 前向分步算法; 曲线下面积; 逻辑斯谛函数; 随机梯度下降

中图分类号: TP311 **文献标志码**: A **DOI:** 10.3969/j.issn.1000-5641.2018.01.009

Forward stagewise additive modeling for entity ranking in documents

WANG Yan-hua

(*School of Data Science and Engineering, East China Normal University, Shanghai* 200062*, China)*

Abstract: Key entities of a document can help to summarize the subjects of the events or the topics that the document describes, which can contribute to applications such as entity-oriented information retrieval and question-answering. However, entities in free text are unordered and hence it is important to rank entities of a document. In this paper, firstly, we make full use of features of entities that extracted from the document and draw support from Wikipedia and Word Embedding to generate external features. Then, we propose a novel ranking model named LA-FSAM(FSAM based on AUC Metric and Logistic Function) which is based on forward stagewise algorithm additive modeling. In LA-FSAM, we employ the AUC(Area Under the Curve) metric to construct the loss function and the logistic function to integrate features of entities. Finally, the stochastic gradient descent is utilized to optimize parameters of LA-FSAM model. After experiments, our evaluation shows the efficiency of the model we proposed.

Key words: entity ranking; forward stagewise additive modeling; area under the curve; logistic function; stochastic gradient descent

收稿日期: 2016-12-01

基金项目: 上海市科技兴农推广项目(2015 第 3-2 号)

作者简介: 王燕华, 男, 硕士研究生, 研究方向为机器学习. E-mail: yhwang917@gmail.com.

0 引 言

随着互联网媒体的兴起, 海量网络新闻文本数据仍在不断增长, 并且这些数据中存储着社会发展过程中重要的舆情事件和热门讨论话题. 当前, 自然语言处理(Natural Language Processing, NLP) 技术的不断成熟使得我们可以利用命名实体识别[1] (Named Entity Recognition, NER) 技术识别文本中出现的实体. 然而, 当文本中存在很多实体时, 如何提取文本中的重要实体来刻画事件(或话题)的发生主体(人物或机构)与发生地点就显得尤为重要, 因为这些实体中仅有一部分是与文本主题相关的, 并且大量的实体可能会淹没或模糊文本主题, 而与文本主题无关的实体可能会偏离事件(或话题)本身并影响对文本主题的理解.

近些年来, 对于文本关键词抽取的研究吸引了越来越多学者的关注[2-11], 它旨在对文本中的词项进行重要性排序, 从而选出 top-k 个词项作为文本的关键词. 其算法主要考虑词项

在文本中的特征, 如词频和位置[2]、文档结构[3]、逆向文档频率[4-5]、词项的主题分布[6-7]等. 尽管存在大量文本关键词抽取的研究工作, 然而对文本重要实体排序的研究却相对欠缺. 文本重要实体不仅可以抽象概括文本主题[12], 而且有助于信息检索[13]和问答系统[14]等方面的研究. 例如, 在信息检索系统中利用实体搜索相关文档时, 在搜索匹配过程中去除文档中的无关实体可以有效地提高文档检索的准确率; 同样在问答系统中, 如何有效地识别文本中的重要实体与无关实体可以更好地求解出问题答案. 因此, 对文本中的实体进行重要性排序具有现实意义.

本文针对单文档的重要实体排序问题, 提出了一种新的排序模型 LA-FSAM (FSAM based on AUC Metric and Logistic Function). LA-FSAM 是一种基于提升(Boosting)的前向分步算法[15], 它充分考虑实体在文档中的重要特性, 提取 4 种实体文档特征, 并利用维基百科和谷歌 Word2Vec 引入 2 种实体外部特征. LA-FSAM 模型运用改进的 AUC 准则构建损失函数, 并使用逻辑斯谛函数混合上述 6 种特征作为基函数, 通过标注训练数据并利用随机梯度下降法学习模型参数, 最终通过实验对比, 证明了本文所提方法的有效性.

本文后续组织如下: 第 1 节简要总结相关研究工作; 第 2 节形式化定义单文档排序问题; 第 3 节详细介绍本文的研究方法; 第 4 节通过实验验证本文所提方法的有效性; 第 5 节总结全文.

1 相关工作

经典的实体排序问题主要应用于信息检索领域, 旨在根据实体与用户搜索的相关性, 对实体进行排序[14,16-18]. 在这些实体排序任务中, 研究的关键在于理解用户搜索并从海量的、异构的数据源中返回与用户搜索最相关的实体集. 文献 [14] 关注于如何理解用户自由无规则的搜索并返回相关的实体或事件, 该文献使用与搜索相关的文档集和知识库来提取相关特征, 分别使用 RankSVM 和 Coordinate Ascent 模型对搜索结果进行相关性排序. 文献[16-17] 通过使用结构化的知识库、实体关系图谱和用户数据来提取有效特征, 并运用基于 GBDT 的模型对搜索实体进行相关性排序. 文献 [18] 通过收集并整合知识库、推特和搜索日志等不同资源(给予权重)的实体描述来提高搜索实体排序的性能, 该方法可实时的增加新的实体描述, 并且通过获取用户如何搜索实体的变化来优化不同资源的实体描述. 学习排序(Learning to Rank, LTR) 方法被广泛应用于信息检索领域并取得了较好的表现, 文献 [14,16-17] 所使用的算法均为学习排序模型. 学习排序方法根据模型的数据输入形式可以分

为基于 Pointwise、Pairwise 和 Listwise 的排序, 这类方法提取特征, 并训练 RankSVM、提升模型或神经网络等排序模型[19]. 因此, 对于文档实体排序问题, 本文考虑提取文档中有关实体的重要特征并通过学习排序方法融合多种特征来实现排序模型.

然而, 本文所提的实体排序问题是指对于任意的文本文档, 根据文本描述内容与主题, 对文本中的实体进行重要性排序, 这与文本关键词抽取的研究有很强的相关性. 文献 [2-5] 利用文本中的词项特征, 例如词性、词频、逆向文档频率、位置和词项长度等, 通过标注训练数据并应用监督学习方法来抽取文本关键词. 非监督学习方法中, 文献 [8] 首次将随机游走算法应用于关键词提取问题, 该文使用词项在滑动窗口内的共同出现来建立词项之间的关系, 随后出现了大量的基于随机游走的关键词抽取研究工作, 他们通过不同的方法建立词项之间的语义关系: 文献 [9] 通过 WordNet 建立词与词之间的语义关系; 文献 [10] 则利用词汇分布表示(Word Embedding) 来计算词项之间的语义关系; 对于其他非监督学习方法, 文献 [7,11] 运用基于图的算法或聚类方法抽取文本关键词. 文本关键词抽取的目标是提取与文本内容和主题相关的词项, 主要考虑名词、动词和形容词, 而文本实体排序则是获取与文本内容和主题相关的重要实体, 他们的研究目标相似, 研究对象略有不同.

基于以上的研究, 本文考虑借助文本关键词抽取的相关研究来提取文本实体特征, 并根据学习排序方法提出 LA-FSAM 模型对文本实体进行排序.

2　单文档实体排序问题

为了更好地介绍 LA-FSAM 算法如何对文本实体进行重要性排序, 本节将详细给出算法涉及的重要概念以及 LA-FSAM 算法的梗概.

定义 1　新闻文档 d_i, 对于任意的新闻文档 d_i, 它由标题 t_i、描述内容c_i和实体列表 $E(d_i)$ 组成, 即 $d_i = \{t_i, c_i, E(d_i)\}$, 其中, 实体列表 $E(d_i) = \{e_{i1}, e_{i2}, \cdots, e_{im}\}$, $e_{ij}(1 \leqslant j \leqslant m)$ 为文档 d_i 中的第 j 个实体, 本文按照实体首次出现的顺序来排序.

定义 2　实体正例集合 P_i 与实体负例集合 N_i, 在任意新闻文档 d_i 的实体列表 $E(d_i)$ 中, 重要的实体构成实体正例集合 P_i, 其他剩余实体构成实体负例集合 N_i.

因此, 根据定义 1 和定义 2 可以将文本重要实体排序问题定义为: 对于任意给定的新闻文档 d_i, 自动地对文档实体列表 $E(d_i)$ 中的实体进行排序, 使得 $E(d_i)$ 中实体正例集合 P_i 中的实体排在实体负例集合 N_i 中的实体的前面.

定义 3　标注新闻文档 dl_i, 对于任意的标注新闻文档 dl_i, 它由标题 t_i、描述内容 c_i、实体正例集合 P_i 与实体负例集合 N_i 构成, 即 $dl_i = \{t_i, c_i, P_i, N_i\}$.

定义 4　标注文档集合 DL, 标注文档集合 DL 由一系列标注新闻文档 dl_i 构成, 即$DL = \{dl_1, dl_2, \cdots, dl_n\}$.

LA-FSAM 是基于前向分步算法的监督学习方法, 该算法主要由以下 4 部分组成.

(1) 命名实体抽取. 对于任意给定的新闻文档 d_i, 本文利用 NER 技术抽取文档标题和描述内容中的所有实体, 借助文献 [20] 所提方法对抽取的实体集合进行归一化处理, 归一化后的实体集合组成该文档重要实体的候选集, 即定义 1 中的实体列表 $E(d_i)$.

(2) 特征提取. 对于任意给定的新闻文档 d_i, 经过第一步命名实体抽取后, 计算实体列表$E(d_i)$中各个实体在 6 种特征下的值来构造特征矩阵 X_i, 对于特征矩阵X_i中的每一种特征做归一化处理, 即特征矩阵X_i中每一列的元素相加之和等于 1. 6 种特征的定义以及计算方法将在第 3 节中详细介绍.

(3) 模型学习. 给定标注文档集合 DL 作为模型输入, LA-FSAM通过不断地叠加新的基函数优化模型, 在每次叠加新的基函数时, 通过极小化损失函数不断地更新模型当前参数, 直至模型收敛. 这里, 模型学习的优化目标为对于任意的新闻文档d_i, 经过排序打分后其实体正例集合P_i中的实体分数高于实体负例集合N_i中的实体分数. 具体的模型算法以及参数学习将在第 3 节中详细介绍.

(4) 实体排序. 通过第(3)部分求解出模型的最优参数后, 任意给定一篇新闻文档 d_i, 经过前两步实体抽取与特征提取后, 通过LA-FSAM模型可以得到该文档中实体的最终分数并对其进行排序.

3 LA-FSAM 算法

本节将详细给出 LA-FSAM 模型算法介绍、基于特征的基函数构造、损失函数构造以及模型参数学习过程.

3.1 LA-FSAM 算法

前向分步算法(FSAM)是一种基于提升(Boosting)的方法, FSAM 方法旨在通过改变训练样本权重, 学习多个基函数, 然后通过将这些基函数进行线性组合提高模型性能[21]. LA-FSAM 模型借鉴 FSAM 算法的思想, 结合文档实体排序问题, 构造适合的基函数与损失函数.特别地, 本文基函数为排序函数, 通过叠加排序函数提高 LA-FSAM 模型性能. 设 $b(X;\overrightarrow{\beta})$ 为排序基函数, $L(P,N,f(X))$ 为损失函数, 其中, X 为实体特征矩阵, $\overrightarrow{\beta}$ 为各个特征$(\overrightarrow{x}_1,\cdots,\overrightarrow{x}_6)$的权重向量, $f(X)$ 为基函数叠加后的加法模型, P 为实体正例集合, N 为实体负例集合, 则 LA-FSAM 算法见算法 1.

算法 1 LA-FSAM 算法

输入: 标注文档集合$DL=\{dl_1,dl_2,\cdots,dl_n\}$, 损失函数$L(P,N,f(X))$, 基函数集$\{b(X;\overrightarrow{\beta})\}$
输出: 加法模型$f(X)$

---

1. 初始化$f_0(X)=0$
2. For $t=1,\cdots,T$

$$(\alpha_t,\overrightarrow{\beta_t})=\arg\min_{\alpha,\overrightarrow{\beta}}\sum_{i=1}^{n}L(P_i,N_i,f_{t-1}(X_i)+\alpha b(X_i;\overrightarrow{\beta}))$$

$$f_t(X)=f_{t-1}+\alpha_t b(X;\overrightarrow{\beta_t})$$

3. 得到加法模型

$$f(X)=\sum_{t-1}^{T}+\alpha_t b(X;\overrightarrow{\beta_t})$$

与前向分步算法思路相似, LA-FSAM算法通过每次只学习一个排序基函数及其参数来不断的优化模型, 不仅求解出高性能的加法模型排序函数$f(X)$, 还降低了优化问题的复杂度.

3.2 基于特征的基函数构造

基函数$b(X;\overrightarrow{\beta})$为排序函数, 给定任意新闻文档 d_i 可以根据实体特征矩阵 X_i 求解出实体列表 $E(d_i)$ 中每个实体的分数. 接下来将分别介绍 6 种实体特征.

(1) 标题频率. 新闻标题作为新闻描述内容的高度概括与总结, 通常会提及重要实体, 例如文本描述事件(或话题)的发生主体(人名, 机构名)或发生地点(位置). 因此, 实体在标题中的出现频率可以作为实体是否重要的判断依据. 任意给定实体列表 $E(d_i)$, 则标题频率特征 $\overrightarrow{x}_1$ 可以表示为

$$\overrightarrow{x}_1 = \Big(\frac{T\text{count}(e_{i1})}{Z_{i1}}, \frac{T\text{count}(e_{i2})}{Z_{i1}}, \cdots, \frac{T\text{count}(e_{im})}{Z_{i1}}\Big)^{\mathrm{T}}, \tag{1}$$

其中, $T\text{count}(e_{ij})$为文档 d_i 中第 j 个实体e_{ij}在标题中出现的次数, Z_{i1}为文档d_i在标题频率特征下的归一化因子.

(2) 描述内容频率. 文本描述内容作为新闻文档的主体内容, 多数实体都来自于这里, 并且通常重要的实体会在描述文本中出现多次. 因此, 特征$\overrightarrow{x}_2$定义为实体在文本描述内容中的频率. 任意给定实体列表$E(d_i)$, 则描述内容频率特征$\overrightarrow{x}_2$可以表示为

$$\overrightarrow{x}_2 = \Big(\frac{C\text{count}(e_{i1})}{Z_{i2}}, \frac{C\text{count}(e_{i2})}{Z_{i2}}, \cdots, \frac{C\text{count}(e_{im})}{Z_{i2}}\Big)^{\mathrm{T}}, \tag{2}$$

其中, $C\text{count}(e_{ij})$为文档 d_i 中第 j 个实体 e_{ij} 在描述内容中出现的次数, Z_{i2} 为文档 d_i 在描述内容频率特征下的归一化因子.

(3) 实体位置. 描述内容文本通常会在文本开头概述新闻事件或话题的主要内容. 因此, 实体在整个文档中首次出现的位置可以作为实体是否重要的判断标准. 任意给定实体列表 $E(d_i)$, 则标题频率特征 $\overrightarrow{x}_2$ 可以表示为

$$\overrightarrow{x}_3 = \frac{1}{Z_{i3}}\Big(\frac{m}{\text{total}_{\text{pos}}}, \frac{m-1}{\text{total}_{\text{pos}}}, \cdots, \frac{1}{\text{total}_{\text{pos}}}\Big)^{\mathrm{T}}, \tag{3}$$

其中, $\text{total}_{\text{pos}} = \sum_{j=1}^{m} j$, Z_{i3}为文档 d_i 在实体位置特征下的归一化因子.

以上 3 个特征为基于实体个体的特征, 而接下来将要介绍的特征为基于实体关系的特征, 通过特定的实体关系构造实体关系矩阵$M_{m\times m}$(m为实体个数), 经过基于 PageRank 的随机游走方法, 并将最终的稳定向量作为实体的特征.

在 PageRank[22]算法中, 设转移矩阵为上文所提实体关系矩阵$M_{m\times m}$, 随机跳转向量$\overrightarrow{t} = (\frac{1}{m}, \frac{1}{m}, \cdots, \frac{1}{m})^{\mathrm{T}}$, 初始概率分布向量$\overrightarrow{v}_0 = (\frac{1}{m}, \frac{1}{m}, \cdots, \frac{1}{m})^{\mathrm{T}}$, γ为随机跳转概率, 则可以通过

$$\overrightarrow{v}_t = (1-\gamma)M\overrightarrow{v}_{t-1} + \gamma\overrightarrow{t} \tag{4}$$

计算得到任意新闻文本 d_i 中各个实体基于某种关系的稳定排序分数值, $t = 1, 2, \cdots$, 直至 $\overrightarrow{v}_t$ 收敛.

(4) 基于实体共现的特征. TextRank[8]首次将基于 PageRank 的随机游走应用于关键词抽取中并取得较好的实验效果, 随后出现了大量基于此方法的关键词提取的研究. TextRank 通过两个实体在滑动窗口内的共同出现建立实体间关系. 本文将滑动窗口定义为每一句话(包括标题). 任意给定实体列表 $E(d_i)$, 对于转移矩阵 M_i 中的任意元素 m_{ijk}, 有

$$m_{ijk} = S\text{count}(e_{ij}, e_{ik}), \tag{5}$$

其中, $S\text{count}(e_{ij}, e_{ik})$为文档$d_i$中实体$e_{ij}$和$e_{ik}$同时出现的句子的个数. 得到转移矩阵$M_i$后对$M_i$做归一化处理.

(5) 基于维基百科的特征. 维基百科页面中的入链接(inlink)和出链接(outlink)提供了实体间丰富的语义关系, 通过这些语义关系可以建立有效的实体关系矩阵M_i. 本文主要使用维基百科页面中的出链接. 任意给定实体列表$E(d_i)$, 对于实体关系矩阵M_i中的任意元素m_{ijk},

$$m_{ijk} = \frac{\text{OutLinkCount}_{e_{ij}}(e_{ik}) + \text{OutLinkCount}_{e_{ik}}(e_{ij})}{2}, \tag{6}$$

其中, $\text{OutLinkCount}_{e_{ij}}(e_{ik})$表示在实体$e_{ij}$的维基百科页面中出链接为实体$e_{ik}$的个数, 同理, $\text{OutLinkCount}_{e_{ik}}(e_{ij})$表示在实体$e_{ik}$的维基百科页面中出链接为实体$e_{ij}$的个数. 得到转移矩阵$M_i$后对$M_i$做归一化处理.

(6) 基于词汇分布表示的特征. 词汇分布表示通过训练海量语料库来得到隐含词汇信息的数值型词向量, 对这些词向量进行计算操作可以得到词汇之间的语义关系, 由于训练词汇分布表示需要海量的语料库并且耗时较长, 本文直接使用谷歌 Word2Vec 作为词汇分布表示并使用余弦相似度来建立任意两个实体之间的语义关系. 任意给定实体列表$E(d_i)$, 对于该实体关系矩阵M_i中的任意元素m_{ijk}, 有

$$m_{ijk} = \frac{\overrightarrow{V}_{eij} \cdot \overrightarrow{V}_{eik}}{|\overrightarrow{V}_{eij}||\overrightarrow{V}_{eik}|}, \tag{7}$$

其中, $\overrightarrow{V}_{eij}$和$\overrightarrow{V}_{eik}$分别表示实体e_{ij}和实体e_{ik}在词汇分布表示中的词向量. 得到转移矩阵M_i后对M_i做归一化处理.

对于任意给定新闻文档, 根据上述方法计算出特征矩阵 X 后, 应用逻辑斯谛函数来结合这 6 种特征, 即

$$b(X; \overrightarrow{\beta}) = \frac{1}{1 + \mathrm{e}^{-X\overrightarrow{\beta}}}, \tag{8}$$

其中, $\overrightarrow{\beta}$为各个特征的权重向量.

3.3 损失函数构造

文献 [23] 提出将 AUC 准则作为排序问题的损失函数, 并且文献 [2] 将 AUC 准则作为关键词排序的损失函数. 因为排序函数旨在通过排序后使实体正例集合中实体的分数尽可能地高于实体负例集合中实体的分数, 这个优化目标和 AUC 准则很相似. 因此, 本文使用 AUC 准则构造损失函数. AUC 的计算定义为

$$\text{AUC} = \frac{\Sigma_{j\in P}\Sigma_{k\in N}\,\text{II}\,(f^j(X) - f^k(X))}{|P||N|}, \tag{9}$$

其中, $f^j(X)$和$f^k(X)$分别表示经函数$f(X)$排序后第j个实体和第k个实体的分数, P为实体正例集合, N为实体负例集合, $\text{II}\,(x)$为指示函数, 有

$$\text{II}\,(x) = \begin{cases} 1, & x > 0, \\ 0, & x \leqslant 0. \end{cases} \tag{10}$$

根据排序函数的优化目标, 定义损失函数$L(P,N,f(X))$为

$$L(P,N,f(X))=\frac{\Sigma_{j\in P}\Sigma_{k\in N}\amalg\left(f^k(X)-f^j(X)\right)}{|P||N|}, \tag{11}$$

但是, 由于本文使用随机梯度下降法极小化损失函数, 根据 AUC 定义的损失函数$L(P,N,f(X))$含有指示函数 $\amalg(x)$, 而指示函数不可导, 因此本文使用 Sigmoid 函数$S(x)$替换指示函数$\amalg(x)$,

$$S(x)=\frac{1}{1+\mathrm{e}^{-x}}. \tag{12}$$

则公式(11)损失函数$L(P,N,f(X))$可以重新定义为

$$L(P,N,f(X))=\frac{\Sigma_{j\in P}\Sigma_{k\in N}S(f^k(X)-f^j(X))}{|P||N|}\ . \tag{13}$$

3.4　参数学习

由第 3.2 节和第 3.3 节可以得到 LA-FSAM 算法中基函数和损失函数, 综合第 3.1 节中 LA-FSAM 算法的介绍, LA-FSAM 算法的最终目标优化函数计算是

$$\min_{\alpha_t,\overrightarrow{\beta}_t}\sum_{i=1}^{n}L\Big(P_i,N_i,\sum_{t=1}^{T}\alpha_t b(X_i;\overrightarrow{\beta}_t)\Big). \tag{14}$$

根据 LA-FSAM 算法, 对于 $t=1,2,\cdots,T$, 通过第 t 次叠加仅学习当前基函数 $b(X;\overrightarrow{\beta}_t)$ 及其系数 α_t, 从而不断逼近目标优化函数(14), 因此, 第 t 次仅需优化损失函数

$$\min_{\alpha_t,\overrightarrow{\beta}}J(\alpha,\overrightarrow{\beta})=\frac{1}{n}\sum_{i=1}^{n}\frac{\sum_{j\in P_i\wedge k\in N_i}S(f^k(X_i)-f^j(X_i))}{|P_i||N_i|}, \tag{15}$$

其中,

$$\begin{cases} f^k(X_i)=f^k_{t-1}(X_i)+\alpha b^k(X_i;\overrightarrow{\beta}), & \text{(16a)}\\ f^j(X_i)=f^j_{t-1}(X_i)+\alpha b^j(X_i;\overrightarrow{\beta}), & \text{(16b)}\end{cases}$$

$b^k(X_i;\overrightarrow{\beta})$ 和 $b^j(X_i;\overrightarrow{\beta})$ 分别表示经基函数$b(X_i;\overrightarrow{\beta})$ 排序后第 k 个和第 j 个实体的分数.

本文使用梯度下降法极小化损失函数(15). 梯度下降法主要有批处理梯度下降法和随机梯度下降法. 考虑到批处理梯度下降法时间复杂度较高, 本文采用随机梯度下降法. 在 LA-FSAM 中, 每次叠加新的基函数时使用随机梯度下降法求解最优参数α 和 $\overrightarrow{\beta}$. LA-FSAM 模型学习算法见算法 2:

算法 2　LA-FSAM 模型学习算法

输入: 标注文档集合$DL=\{dl_1,dl_2,\cdots,dl_n\}$, 学习速度 lr
输出: α和$\overrightarrow{\beta}$

---

1. $j=0$
2. 初始化$\alpha^{(0)}$, $\overrightarrow{\beta}^{(0)}$

3. 若$J(\alpha, \overrightarrow{\beta})$未收敛, 则循环做

对于训练数据集中的每一篇文档 dl_i

$\alpha^{(j+1)} = \alpha^{(j)} - lr\frac{\partial Ji(\alpha^{(j)}, \overrightarrow{\beta}^{(j)})}{\partial \alpha}$

$\overrightarrow{\beta}^{(j+1)} = \overrightarrow{\beta}^{(j)} - lr\frac{\partial Ji(\alpha^{(j)}, \overrightarrow{\beta}^{(j)})}{\alpha \overrightarrow{\beta}}$

$j = j + 1$

算法 2 中$J_i(\alpha, \overrightarrow{\beta})$为基于文档$dl_i$的损失函数

$$J_i(\alpha, \overrightarrow{\beta}) = \frac{\sum_{j\in P_i \wedge k \in N_i} S(f^k(X_i) - f^j(X_i))}{|P_i||N_i|}. \tag{16}$$

下面给出算法2中$\frac{\partial J_i(\alpha, \overrightarrow{\beta})}{\partial \alpha}$和$\frac{\partial J_i(\alpha, \overrightarrow{\beta})}{\alpha \overrightarrow{\beta}}$的计算方法.

$$\frac{\partial J_i(\alpha, \overrightarrow{\beta})}{\partial \alpha} = \frac{\sum_{j\in P_i \wedge k \in N_i} \frac{\partial S(\Delta f^{kj})}{\partial \Delta f^{kj}} \left(\frac{\partial f^k(X_i)}{\partial \alpha} - \frac{\partial f^j(X_i)}{\partial \alpha} \right)}{|P_i||N_i|}, \tag{17}$$

$$\frac{\partial J_i(\alpha, \overrightarrow{\beta})}{\alpha \overrightarrow{\beta}} = \frac{\sum_{j\in P_i \wedge k \in N_i} \frac{\partial S(\Delta f^{kj})}{\partial \Delta f^{kj}} \left(\frac{\partial f^k(X_i)}{\partial \overrightarrow{\beta}} - \frac{\partial f^j(X_i)}{\partial \overrightarrow{\beta}} \right)}{|P_i||N_i|}, \tag{18}$$

其中,

$$\begin{cases} \Delta f^{kj} = f^k(X_i) - f^j(X_i), & \text{(20a)} \\ \dfrac{\partial S(\Delta f^{kj})}{\partial \Delta f^{kj}} = S(\Delta f^{kj}) \cdot (1 - S(\Delta f^{kj})). & \text{(20b)} \end{cases}$$

可以根据公式(16)可以得到

$$\begin{cases} \dfrac{\partial f^k(X_i)}{\partial \alpha} = b^k(X_i; \overrightarrow{\beta}), & \text{(21a)} \\ \dfrac{\partial f^j(X_i)}{\partial \alpha} = b^j(X_i; \overrightarrow{\beta}), & \text{(21b)} \end{cases}$$

$$\begin{cases} \dfrac{\partial f^k(X_i)}{\partial \overrightarrow{\beta}} = \alpha \cdot b^k(X_i; \overrightarrow{\beta}) \cdot (1 - b^k(X_i; \overrightarrow{\beta}))(X_i^k)^{\mathrm{T}}, & \text{(22a)} \\ \dfrac{\partial f^j(X_i)}{\partial \overrightarrow{\beta}} = \alpha \cdot b^j(X_i; \overrightarrow{\beta}) \cdot (1 - b^j(X_i; \overrightarrow{\beta}))(X_i^j)^{\mathrm{T}}, & \text{(22b)} \end{cases}$$

其中, X_i^k 和 X_i^j 分别表示矩阵 X_i 的第 k 行和第 j 行.

将公式(20)、(21)、(22)分别代入到公式(18)、(19)可以求出算法 2 中的$\frac{\partial J_i(\alpha, \overrightarrow{\beta})}{\partial \alpha}$和$\frac{\partial J_i(\alpha, \overrightarrow{\beta})}{\alpha \overrightarrow{\beta}}$.

4 实 验

本节将通过系统的实验来评测 LA-FSAM 模型的表现. 下文将分别对数据集、相关参数设置、基函数叠加数量对 LA-FSAM 模型的影响、基线方法与 LA-FSAM 的对比以及案例进行详细说明.

4.1 数据集介绍

目前没有有关文本实体排序相关的公开测试数据集, 本文使用来自于文献 [24] 的公开新闻报道数据集, 并对其进行人工标注当作评测数据. 该数据集中的文本主要来着谷歌搜索中中东地区冲突事件及话题报道, 这里仅使用每篇报道的标题与描述内容. 本文从数据集中随机选取 500 条新闻报道, 使用 Standford-NLP[1]工具抽取文本实体并人工对其每篇文章进行重要实体标注, 最后将 80% 的标注数据当作训练数据, 剩余 20% 的标注数据用做测试数据. 数据集中重要实体数量统计如表 1 所示.

表 1　数据集重要实体数量统计

Tab. 1　Count of key entities statistics of documents

实体数量	文档数量	占文档比例/%
$\geqslant 1$	500	100
$\geqslant 2$	487	97.4
$\geqslant 3$	389	77.8
$\geqslant 4$	201	40.2
$\geqslant 5$	89	17.8

4.2　相关参数设置

参数lr: 参数 lr 为学习速度, 较大的学习速率可以加快算法 2 中优化速度, 但是过大可能会出现震荡无法收敛, 经过实验本文选取 $lr = 0.01$.

参数 γ: 参数 γ 为随机跳转概率, 在 $0.1\leqslant \gamma \leqslant 0.9$ 的范围内, 本文以平均损失函数值与平均 AUC 值为标准, 通过对照实验来分析γ对 LA-FSAM 模型的影响. 由图1可以看出当 $\gamma \geqslant 0.3$ 时平均损失函数值较小、平均AUC值较大, 综合考虑平均损失函数值与平均 AUC 值随 γ 的变化, 当 $0.7\leqslant \gamma \leqslant 0.9$ 时平均损失函数值与平均 AUC 值趋于平稳, 因此本文选取 $\gamma = 0.8$.

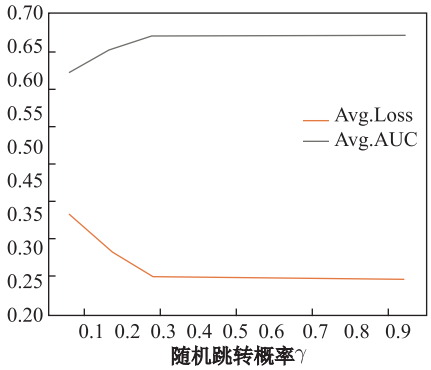

图 1　随机跳转概率 γ 对 LA-FASM 的影响

Fig. 1　Evaluation of γ

4.3　基函数叠加数量对 LA-FSAM 模型的影响

由第 3.1 节算法 1 可以知道 LA-FSAM 算法通过不断累加新的排序基函数来优化模型. 对于模型基函数数量 t, 根据图 2(横坐标为基函数累加数量, 纵坐标为平均损失函数值与平均AUC值)可知, 随着基函数数量的增加平均损失函数值不断减小, 平均 AUC 值不断增大, 并且当 $t \leqslant 10$ 时随着t的增加平均损失函数值与平均 AUC 值变化明显, 当 $t > 30$ 时平均损失函数与平均 AUC 值逐渐趋于平稳. 因此, 基于前向分步算法的 LA-FSAM 模型对实体排序有效.

4.4　不同特征 LA-FSAM 模型的影响

为了衡量不同特征对于 LA-FSAM 模型的影响, 对于第 3.2 节介绍的每一个特征, 本实验环

节通过分别移除该特征, 使用剩余特征训练模型, 分析在缺少该特征情况下 LA-FSAM 的排序性能. 同样地, 以平均损失函数值与平均 AUC 值作为评价指标, 由图3可以看出, 特征 3 实体位置和特征 1 标题频率对 LA-FSAM 排序模型影响最大.

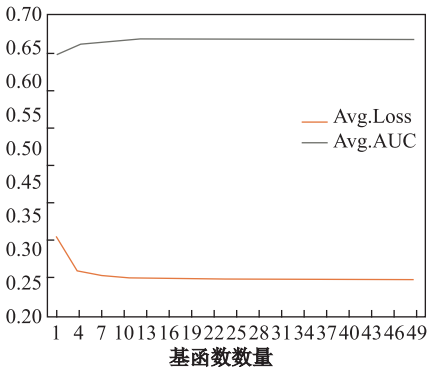

图 2 基函数数量对平均损失函数与平均AUC的影响

Fig. 2 Evaluation of number of basis function

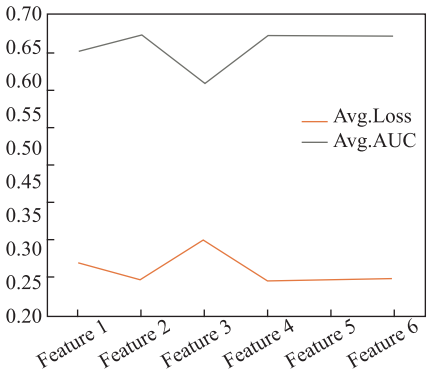

图 3 不同特征对LA-FASM的影响

Fig. 3 Evaluation of features

4.5 基线方法介绍

为了证明 LA-FSAM 模型对文本实体排序的有效性, 本文选取 6 种方法作为基线方法, 它们分别如下.

基于频率的统计(Frequency): 此方法仅考虑实体在文档中的出现频率.

基于位置的统计(Position): 此方法仅考虑实体在文档中首次出现的位置, 并且随着首次出现位置的不断靠后重要程度依次降低.

基于共同出现的随机游走(TextRank): 此方法主要基于文献 [8] 所提, 通过实体在滑动窗口内的共同出现建立实体间语义关系, 最后通过随机游走对实体节点进行排序, 这里滑动窗口设置为文本中的每个句子.

基于维基的随机游走(WikiRank): 这个方法与 TextRank 类似, 但是 WikiRank 通过维基百科链接(inlinks and outlinks)建立实体语义关系.

基于词汇分布表示的随机游走(WordEmbeddingRank): 此方法主要基于文献 [10] 所提, 通过借助词汇分布表示获得蕴含语义的词汇向量, 并通过向量间的相似度建立实体间语义关系, 该算法同样使用随机游走对实体结点进行排序.

RankSVM: RankSVM[25]是基于 Pairwise 的学习排序算法, 并且 RankSVM 也是目前最先进的排序算法之一, 被广泛应用于排序问题中.

4.6　实验结果

根据表 1 可以看出少于 20% 的文档集含有超过5个的重要实体. 因此, 本文使用 LA-FSAM 以及第 4.5 节所述 6 种基线方法, 对测试数据进行实体排序, 并分别对排序后的结果做 top-1, top-2, top-3, top-4 和平均 AUC 值的实验对比, 实验对比结果如表 2 所示. 在上述基线方法中, 基于位置的统计方法和 RankSVM 取得了较好的实验结果; 但是, 与 6 种基线方法对比, 本文 LA-FSAM 模型取得最优的表现.

表 2　LA-FSAM 模型与其他算法平均准确率对比

Tab. 2　Comparison of average precision between LA-FSAM and other methods

Method	Avg.P@1/%	Avg.P@2/%	Avg.P@3/%	Avg.P@4/%	Avg.AUC/%
Frequency	86.4	82.8	81.1	78.6	69.2
Position	92.2	82.5	72.2	63.3	86.2
TextRank	83.5	76.3	67.5	69.0	65.5
WikiRank	68.0	62.6	61.4	68.5	54.5
WordEmbeddingRank	55.3	52.0	57.0	61.9	61.9
RankSVM	92.2	85.9	84.6	77.3	87.6
LA-FSAM	**94.2**	**88.9**	**85.5**	**79.8**	**88.3**

4.7　案例分析

本节从测试集中随机抽取一篇文章做案例分析. 文章标题和文章内容如下所示(由于篇幅较长, 文中较多内容用"……"替代).

U.N. says ***Syria*** death toll has likely surpassed 100,000

An image provided by opposition activists in April shows a mass grave said…… (***Aleppo Media*** Center/Associated……) ***BEIRUT*** – The ever-escalating death toll in ***Syria***'s two-year civil war has likely surpassed 100,000, the ***United Nations*** said Thursday, …… carried out for the ***U.N.*** human rights office. …… according to the ***U.N.*** analysis. …… over the past year," ***U.N.*** High Commissioner for Human Rights ***Navi Pillay*** said ……, including babies, being massacred," ***Pillay*** said. The death count — which likely underestimates the total, ***Pillay*** said — ……, the ***U.N.*** noted, ……more deaths. But ***Pillay*** noted that ……the strategic city of ***Qusair***, ……the northern city of ***Aleppo***, which …… for almost a year. Any battle to retake ***Aleppo*** would likely be protracted and bloody. The ***U.N.*** has repeatedly ……have failed. A joint ***U.S.***-Russian initiative …… Syrian President ***Bashar Assad*** agree to step down, ……Russian brokered peace talks.

该报道文主要针对"联合国(U.N.)高级人权专家皮莱(Navi Pillay)发表的叙利亚(Syria)内战所引发的人员伤亡分析声明"做报道. 报道中抽取的实体已在文本中用加粗斜体标记; 人工标注重要实体有: U.N., Syria, Navi Pillay; 经 LA-FSAM 排序后实体分数从高到低依次为: U.N., Syria, Navi Pillay, Aleppo Media, BEIRUT, Qusair, Aleppo, U. S., Bashar Assad.

5　总　　结

文本重要实体不仅可以抽象概括文本描述事件或话题的发生主体与发生位置, 还有助于信息检索和问答系统等方面的研究, 因此, 对文本实体进行排序具有现实价值. 本文针对该问题提出了一种基于前向分步算法的排序模型LA-FSAM, 该模型使用逻辑斯谛函数混合六种有效特

征, 运用改进的 AUC 准则构建损失函数, 通过标注训练数据并利用随机梯度下降法学习模型参数. 最终通过与其他基线方法的实验对比证明了 LA-FSAM 方法的有效性.

[参 考 文 献]

[1] FiNKEL J R, GRENAGER T, MANNING C. Incorporating non-local information into information extraction systems by gibbs sampling [C]//Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2005: 363-370.

[2] ZHANG W, FENG W, WANG J Y. Integrating semantic relatedness and words' intrinsic features for keyword extraction[C]//Proceedings of the 23rd International Join Conference on Artificial Intelligence. 2013: 2225-2231.

[3] HOFMANN K, TSAGKIAS M, MEIJ E, et al. The impact of document structure on keyphrase extraction[C]//Proceedings of the 18th ACM conference on Information and knowledge management. ACM, 2009: 1725-1728.

[4] LI Z H, ZHOU D, JUAN Y F, et al. Keyword extraction for social snippets[C]//Proceedings of the 19th International Conference on World Wide Web. ACM, 2010: 1143-1144.

[5] JIANG X, HU Y H, LI H. A ranking approach to keyphrase extraction[C]//Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2009: 756-757.

[6] ZHANG F, HUANG L E, PENG B. WordTopic-MultiRank: A new method for automatic keyphrase extraction[C]// Proceedings of the 6th International Joint Conference on Natural Language. ACL, 2013: 10-18.

[7] LIU Z Y, HUANG W Y, ZHENG Y B, et al. Automatic keyphrase extraction via topic decomposition[C]//Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2010: 366-376.

[8] MIHALCEA R, TARAU P. TextRank: Bringing order into texts[C]//Conference on Empirical Methods in Natural Language Processing. ACL, 2004: 404-411.

[9] WANG J H, LIU J Y, WANG C. Keyword extraction based on pagerank[C]//Pacific-Asia Conference on Knowledge Discovery and Data Mining. Berlin: Springer, 2007: 857-864.

[10] WANG R, LIU W, MCDONALD C. Using word embeddings to enhance keyword identification for scientific publications [C]//Australasian Database Conference. Berlin: Springer International Publishing, 2015: 257-268.

[11] LIU Z Y, LI P, ZHENG Y B, et al. Clustering to find exemplar terms for keyphrase extraction[C]//Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1. Association for Computational Linguistics, 2009: 257-266.

[12] DEMARTINI G, MISSEN M M S, BLANCO R, et al. Entity summarization of news articles[C]//Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2010: 795-796.

[13] BASHIR S, AFZAL W, BAIG A R. Opinion-based entity ranking using learning to rank[J]. Applied Soft Computing, 2016, 38: 151-163.

[14] SCHUHMACHER M, DIETZ L, PONZETTO S P. Ranking entities for Web queries through text and knowledge[C]//Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. ACM, 2015: 1461-1470.

[15] HASTIE T, FRIEDMAN J ,TIBSHIRANI R. The Elements of Statistical Learning[M]//Springer Series in Statistics. New York: Springer-Verlag ,2001: 342-343.

[16] KANG C S, YIN D W, ZHANG R Q, et al. Learning to rank related entities in Web search[J]. Neurocomputing, 2015, 166: 309-318.

[17] KANG C S, VADREVU S, ZHANG R Q, et al. Ranking related entities for Web search queries[C]//Proceedings of the 20th International Conference Companion on World Wide Web. ACM, 2011: 67-68.

[18] GRAUS D, TSAGKIAS M, WEERKAMP W, et al. Dynamic collective entity representations for entity ranking[C]//Proceedings of the 9th ACM International Conference on Web Search and Data Mining. ACM, 2016: 595-604.

[19] LI H. Learning to Rank for Information Retrieval and Natural Language Processing[C/OL]//Synthesis Lectures on Human Language Technologies #26. 2nd ed. [S.l]: Morgan and Claypool Publishers, 2014[2016-07-01]. http://www.morganclaypool.com/doi/ suppl/10.2200/S00607ED2V01Y201410HLT026/suppl_file/li_Ch1.pdf.

[20] JIJKOUN V, KHALID M A, MARX M, et al. Named entity normalization in user generated content[C]//Proceedings of the 2nd Workshop on Analytics for Noisy Unstructured Text Data. ACM, 2008: 23-30.

[21] 李航. 统计学习方法[M]. 北京: 清华大学出版社, 2012: 137-145.

(下转第 145 页)

[19] 关键, 张晓利, 简涛, 等. 分布式目标的子空间双门限 GLRT-CFAR 检测 [J]. 电子学报, 2012, 9: 1759-1764.
[20] 陈建军, 黄孟俊, 赵宏钟, 等. 相参雷达时频域 CFAR 检测门限获取方法研究 [J]. 电子学报, 2013, 8: 1634-1638.
[21] GURAKAN B, CANDAN C, CILOGLU T. CFAR processing with switching exponential smoothers for nonhomogeneous environments [J]. Digital Signal Processing, 2012, 22: 407-416.
[22] WEINBERG G V. Management of interference in Pareto CFAR processes using adaptive test cell analysis [J]. Signal Processing, 2014, 104: 264-273.

(责任编辑: 李　艺)

(上接第 102 页)

[22] BRODER A, KUMAR R, MAGHOUL F, et al. Graph structure in the Web[J]. Computer Networks, 2000, 33(1): 309-320.
[23] FENG W, WANG J Y. Incorporating heterogeneous information for personalized tag recommendation in social tagging systems[C]// Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2012: 1276-1284.
[24] TRAN G, ALRIFAI M, HERDER E. Timeline summarization from relevant headlines[C]//European Conference on Information Retrieval. Springer International Publishing, 2015: 245-256.
[25] JOACHIMS T. Training linear SVMs in linear time [C]//Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2006: 217-226.

(责任编辑: 李　艺)

(上接第 134 页)

[8] EARLE P S, SHEARER P M. Characterization of global seismograms using an automatic-picking algorithm [J]. Bulletin of the Seismological Society of America, 1994, 84(2): 366-376.
[9] MAEDA N. A method for reading and checking phase times in auto-processing system of seismic wave data [J]. Zisin, 1985, 38(3): 365-379.
[10] MARQUARDT D W. An algorithm for least-squares estimation of nonlinear parameters [J]. Journal of the Society for Industrial and Applied Mathematics, 1963, 11(2): 431-441.

(责任编辑: 李　艺)