多粒度关系融合的微博信念网络检索模型

张雄涛，甘明鑫，李　硕
北京科技大学 经济管理学院，北京 100083

摘要：微博检索系统是微博平台实现个性化信息过滤的重要工具，建立合理的微博检索模型不仅有利于满足用户的个性化信息需求，而且有利于提升微博平台的信息服务水平。与传统文本检索相比，微博检索面临两方面挑战，一是较短的微博查询语句难以准确表达用户的检索需求，二是较短的微博文档难以充分表现微博的语义特征。这使查询语句与文档之间难以准确进行匹配计算，致使微博检索性能受限。

结合微博术语和文档的特点，将术语关系和文档关系融入信念网络检索模型，提出多粒度关系融合的微博信念网络检索模型。首先，基于混合语义信息和时间信息对微博术语关系进行量化，以更准确地建立微博术语之间的相关性；其次，基于混合语义信息和作者信息对微博文档关系进行量化，以更准确地建立微博文档之间的相关性；最后，结合量化的术语关系和文档关系，在基本信念网络检索模型的基础上，给出微博信念网络检索模型的概率推导过程。采用网络爬虫工具，从新浪微博平台获取真实的微博数据，对微博信念网络检索模型的有效性和合理性进行验证。

研究结果表明，微博术语之间存在语义相关性和时间相关性；微博文档之间不仅存在语义相关性，还存在作者信息相关性；与主流的微博检索模型相比，微博信念网络检索模型在多项信息检索指标上体现出较优的检索性能。消融实验结果表明，微博术语之间存在语义相关性和时间相关性；微博文档之间不仅存在语义相关性，还存在作者信息相关性；与仅考虑单一粒度关系或不考虑任何关系的信念网络检索模型相比，融合多粒度关系的信念网络检索模型体现出较优性能。

聚焦于微博检索情景，综合查询语句扩展和文档扩展的优势实现微博检索，有效克服了微博检索面临的挑战，显著提升了微博检索的性能。在信息过载的背景下，微博信念网络检索模型不仅有助于进一步提升社交媒体平台的信息服务水平，也可以为社交媒体平台开发合理的检索系统提供借鉴。

关键词：多粒度关系融合；微博检索；信念网络检索模型；术语关系；文档关系
中图分类号：G203　　**文献标识码**：A　　**doi**：10.3969/j.issn.1672－0334.2022.05.005
文章编号：1672－0334(2022)05－0067－13

收稿日期：2021－08－27　　**修返日期**：2022－05－10
基金项目：国家自然科学基金 (72271024, 71871019)

作者简介：张雄涛，北京科技大学经济管理学院博士研究生，研究方向为社交媒体计算、信息检索与推荐等，代表性学术成果为“Integrating community interest and neighbor semantic for microblog recommendation”，发表在 2021 年第 2 期《International Journal of Web Services Research》，E-mail：B20190412@xs.ustb.edu.cn
甘明鑫，管理学博士，北京科技大学经济管理学院教授，研究方向为社交媒体计算和推荐系统等，代表性学术成果为“A knowledge-enhanced contextual bandit approach for personalized recommendation in dynamic domain”，发表在 2022 年第 251 卷《Knowledge-Based Systems》，E-mail：ganmx@ustb.edu.cn
李硕，北京科技大学经济管理学院硕士研究生，研究方向为社交媒体计算和智能数据分析等，E-mail：18511827966@163.com

引言

社交媒体快速发展的背景下，微博已成为人们社会交往、信息获取和分享的重要平台[1]。海量微博信息使用户面临不可避免的信息过载难题，即用户难以从大量的微博信息中快速、准确地获取感兴趣且有价值的信息[2]。微博检索可以借助信息检索技术为用户生成个性化微博信息[3]，故相关研究被广泛关注。

与传统文本检索相比，微博检索主要面临两方面挑战，一是较短的查询语句 (query) 使用户的检索需求难以被准确表达[4]，二是较短的文档 (document) 限制了微博的语义表达[5]，导致查询语句与文档无法有效实现匹配计算，即词不匹配[6]。大量研究表明，查询扩展和文档扩展可以通过引入有效特征来修正或完善查询语句或文档的表示[4]。因此，查询扩展和文档扩展是解决微博检索中词不匹配问题的重要途径。信念网络检索模型是一种基于贝叶斯网络的经典信息检索模型[7]，通过将术语关系或文档关系融入概率推导过程，可以使模型有效发挥查询扩展或文档扩展的优势[8]。鉴于此，本研究结合微博术语和文档的特点，对术语之间和文档之间的关系进行量化，并将其融入信念网络检索模型，提出多粒度关系融合的微博信念网络检索模型。

1 相关研究评述

微博检索，又称微博搜索，是微博平台为用户实现个性化信息查询的一种重要的决策支持系统[9]。依据微博检索的特点，相关研究主要从查询建模、文档建模和查询－文档匹配建模3个视角展开。结合本研究目标，分别对微博检索中的查询扩展研究、文档扩展研究和查询－文档匹配研究进行梳理。

1.1 面向微博检索的查询扩展

面向微博检索的查询扩展研究主要结合查询语句的上下文信息获取查询扩展词，修正或扩展原始查询语句，以更充分地实现微博查询建模。为从查询语句上下文中获取高质量的查询扩展词，ZINGLA et al.[4] 提出混合的查询扩展方法，将查询扩展过程分为候选词生成和选择两个阶段，并通过设计结合外部资源的关联规则实现查询扩展。为缓解Twitter检索中的词不匹配问题，NASIR et al.[10] 使用来自Freebase的知识术语对原始查询语句进行扩展，并在扩展方法中融入时间证据动态调整检索结果，以实时地反映用户的个性化检索意图。上述两项研究中的查询扩展词依赖于人工更新的外源知识库，对于网络新词层出不穷的微博环境难以发挥理想性能。为结合用户信息更有效地实现个性化微博检索，吴树芳等[11] 提出一种多源信息融合的微博检索方法，该方法混合全集微博数据、微博作者信息和微博本身信息，在统计语言模型的基础上对微博文档的语义稀疏性进行有效缓解。考虑微博词汇在时间层面的相关性，韩中元等[12] 基于查询词汇的时间上下文获取查询扩展词，并结合查询扩展词在风险最小化模型的基础上实现微博检索。与其他研究相比，利用词汇的时间相关性可以有效弥补外源知识无法捕获的术语时间关联，但该研究忽略了术语之间的语义关系，因此无法避免时间相关性对查询扩展带来的噪声信息。区别于已有研究，本研究综合微博术语的语义信息和时间信息量化微博术语之间的关系，并结合量化的术语关系在信念网络模型的基础上实现微博查询扩展，以通过更有效地建模微博查询表示进一步提升微博检索性能。

1.2 面向微博检索的文档扩展

微博文档属于典型的社交媒体文本，具有情感极性显著[13−14]、特征稀疏性[15] 和多模态性[16] 等显著特点。面向微博检索的文档扩展研究主要通过利用丰富的微博上下文对微博短文本进行扩展，以更充分地实现微博文档建模[17]。考虑到微博具有的短文本特性，ZHU et al.[9] 和KALLOUBI et al.[18] 将微博文本视作查询语句，采用伪相关反馈方法在保证微博文本的网络化特性的前提下，将伪相关文档作为扩展语料对原始微博文档进行扩展，有效提升了微博检索性能；SAMUEL et al.[19] 结合Twitter和Facebook上的时空上下文信息对社交媒体文本进行语义扩展，不仅对微博文档的及时性特征进行合理建模，而且有效地将空间特征融入微博文档的表示，在微博事件检测中表现优异；寇菲菲等[20] 针对主题模型应用于微博短文本语义建模时存在的不足，提出面向搜索的微博短文本语义建模方法，以具有局部语义的词向量为基础，通过计算单词之间相似度对微博短文本进行扩展，以此缓解短文本的语义稀疏性，并实现局部语义与全局语义的相互补充。已有研究虽已取得一定进展，但主要通过丰富微博文档的语义特征实现微博检索，忽略了微博文档的作者属性。由于具有相似画像的微博用户倾向于发布相似主题的微博，故微博作者的同质性可为微博文档之间的相似性提供依据[21]。基于已有研究的不足，本研究综合微博文档的语义信息和作者信息量化微博文档之间的关系，并结合量化的文档关系在信念网络模型的基础上实现微博文档扩展，以通过更有效地建模微博文档表示进一步提升微博检索性能。

1.3 面向微博检索的查询语句与文档的匹配

与查询扩展和文档扩展的相关研究相比，面向微博检索的查询语句与文档的匹配研究更关注于两者之间的交互[22]，主要结合用户的检索需求优化查询语句与文档之间的匹配过程。针对用户对特定主题微博的检索需求差异，ZHOU et al.[23] 提出基于深度强化学习的安全话题搜索方法，将特定话题的搜索建模为一个连续状态的马尔可夫决策过程，并通过动态组合卷积神经网络和长短时记忆网络评估查询语句与微博文档的相关性，克服了仅依据内容相似性建模匹配的不足。考虑到多模态情景对社交媒体中人物检索的显著影响，TIAN et al.[24] 融合深度多模态分析、跨模态关联学习和多模态信息挖掘，提出一种新的深度跨模态信息检索方法，有效提升了跨模态

信息匹配的准确性; 吴树芳等[25]考虑到用户检索需求受社交因素的影响, 利用微博作者之间的社交关系对查询似然模型中文档语言模型的估计进行改进, 进一步提升了查询似然模型的社交化检索性能。综上所述, 已有关于查询语句与文档匹配的研究主要在查询语句和文档被充分表示的前提下实现, 忽略了较短的微博查询语句和微博文档对微博检索造成的消极影响。区别于已有研究, 本研究在信念网络模型的基础上, 将微博术语关系和文档关系融入查询语句与文档的匹配计算, 以通过对微博查询和微博文档进行合理的建模进一步提升微博检索的性能。

2 研究基础

2.1 基本信念网络检索模型

基本信念网络检索模型是基于贝叶斯网络的经典信息检索模型, 可以依据查询语句和文档建立网络拓扑结构, 并通过合理设计概率推导过程建立查询语句与文档之间的关系[26]。图 1 给出基本信念网络检索模型的拓扑结构, 基本信念网络主要包括 3 类节点集, 即查询节点集Q、术语节点集U和文档节点集D。其中, q 为查询语句; k 为术语, i 为术语序号, t 为术语数量, $i=1,\cdots,t$; d 为文档, j 为文档序号, n 为文档数量, $j=1,\cdots,n$。若 k_i 术语是 q 查询语句的一个查询术语, 则有一条有向边从 k_i 指向q; 若 k_i 术语是 d_j 文档的一个索引术语, 则有一条有向边从 k_i 指向 d_j。

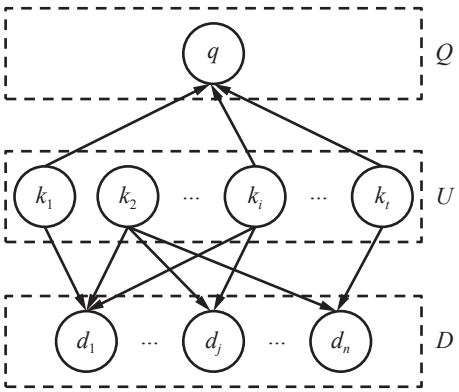


图 1 基本信念网络检索模型的拓扑结构

Figure 1 Topological Structure of the Basic Belief Network Retrieval Model

在基本信念网络检索模型中, 所有术语组成一个概念空间, 令 u 概念表示概念空间中的一个子集, 则 q 和 d_j 均可被视为概念空间中的一个概念, 因此信息检索过程可被视为 d_j 与 q 的概念匹配过程[7], 计算匹配的公式为

$$P(d_j|q)=\eta\sum_{\forall u}P(d_j|u)P(q|u)P(u) \tag{1}$$

其中, $P(d_j|q)$为d_j与 q 的匹配度; η为规范化常数; $P(d_j|u)$为d_j与 u 的匹配度; $P(q|u)$为 q 与 u 的匹配度; $P(u)$为u 的先验概率, 通常假设其符合均匀分布。不同检索策略中$P(d_j|u)$和$P(q|u)$对应不同的计算方式[8], 例如:

当$P(d_j|u)$的计算方式被规定为

$$P(d_j|u)=\frac{\sum_{i=1}^{t}w_{k_i,d_j}w_{k_i,u}}{\sqrt{\sum_{i=1}^{t}w_{k_i,d_j}^2}\sqrt{\sum_{i=1}^{t}w_{k_i,u}^2}} \tag{2}$$

其中, w_{k_i,d_j}为 k_i 在 d_j 中的权重, $w_{k_i,u}$为 k_i 在 u 中的权重。

当$P(q|u)$的计算方式被规定为

$$P(q|u)=\begin{cases}1,\varphi(q,u)\neq 0\\ 0,\varphi(q,u)=0\end{cases} \tag{3}$$

其中, $\varphi(q,u)$为函数, 被定义为q 与u 交集包含的元素个数。

此时, 信念网络检索模型可以转化为向量空间模型[27]。

2.2 基本信念网络检索模型的不足分析

由于基本信念网络检索模型假设术语之间和文档之间是相互独立的[28], 因此只有保证术语同时出现在查询语句和文档中, 基本信念网络才可以发挥作用。然而, 在微博检索情景下, 长度较短的微博查询语句和微博文档包含的有效关键词有限, 导致基本信念网络检索模型在微博检索任务中面临严重的词不匹配问题[6], 即仅依据关键词匹配难以准确实现信息检索功能。图 2 给出一个微博检索系统示例, 实线箭头表示术语与文档之间的关键词匹配关系, 虚线箭头表示术语之间或文档之间的相关关系。在此微博检索系统中, 假如用户发出的查询术语为“国庆节”, 依据基本的信念网络检索模型仅能检索到“微博文档 3”。然而, 结合实际情况可以了解到, 微博文档 1、微博文档 2、微博文档 4、微博文档 5 均属于查询术语“国庆节”的相关文档, 但由于基本信念网络检索模型仅能依据关键词匹配实现检索, 故无法充分发挥检索性能。

作为“国庆节”的相关术语, “十一”和“黄金周”的引入可以有效丰富用户的检索需求。因此, 如果在“国庆节”的基础上融入“十一”和“黄金周”等相关术语, 可以使微博检索性能得以有效提升。此外, 作为微博文档 3 的相关文档, 微博文档 4 和微博文档 5 的引入可以有效缓解微博文档 3 的语义稀疏性。因此, 如果在微博文档 3 的基础上融入微博文档 4 和微博文档 5, 同样可以使微博检索性能得以进一步提升。由此可见, 融合术语关系和文档关系理论上可以有效避免基本信念网络检索模型存在的缺陷, 即融合术语关系可以有效完善用户的检索需求, 融合文档关系可以有效避免微博文档固有的语义稀疏性。

3 微博信念网络检索模型

结合基本信念网络检索模型在微博检索任务中

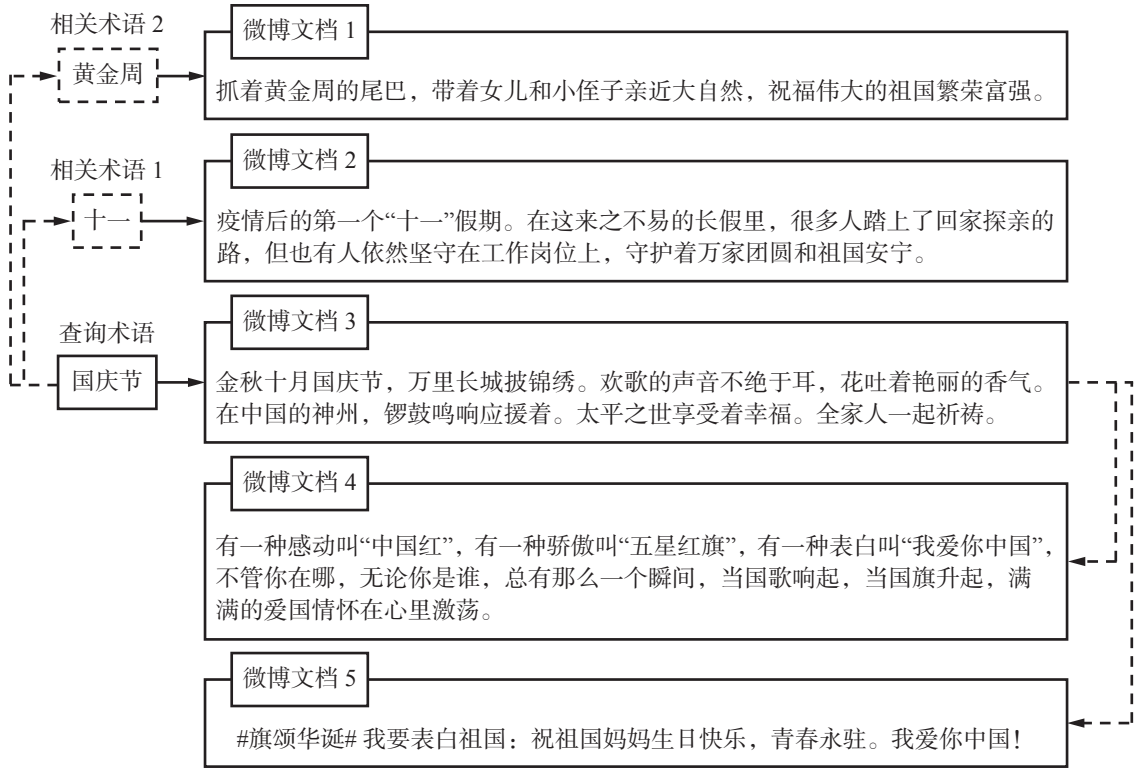


图 2 微博检索系统示例

Figure 2 An Example of the Microblog Retrieval System

存在的不足，本研究在基本信念网络检索模型的基础上，通过融合术语关系和文档关系提出微博信念网络检索模型。首先，介绍微博信念网络检索模型的拓扑结构；其次，分别对微博术语关系和微博文档关系进行量化；最后，结合量化的术语之间关系和文档之间关系给出微博信念网络检索模型的概率推导过程。

3.1 拓扑结构

图 3 为微博信念网络检索模型的拓扑结构，与图 1 的基本模型拓扑结构相比，微博信念网络检索模型考虑了微博术语关系和微博文档关系。图 3 中，微博信念网络检索模型自上而下有 5 类节点和 4 类关系，5 类节点包括查询语句节点集 Q、扩展术语节点集 U'、术语节点集 U、文档节点集 D 和扩展文档集 D'，4 类关系包括查询语句 − 术语关系、术语 − 术语关系、术语 − 文档关系和文档 − 文档关系。

(1) Q 由一个查询语句 q 组成，$Q=\{q\}$；

(2) U 由 t 个索引术语 k 组成，$U=\{k_1,\cdots,k_i,\cdots,k_t\}$；

(3) U' 由 U 复制而成，$U'=\{k'_1,\cdots,k'_i,\cdots,k'_t\}$，由于 U' 与 U 包含相同的术语节点，因此 U' 与 U 互为包含关系；

(4) D 由 n 个文档 d 组成，$D=\{d_1,\cdots,d_j,\cdots,d_n\}$；

(5) D' 由 D 复制而成，$D'=\{d'_1,\cdots,d'_j,\cdots,d'_n\}$，由于 D' 与 D 包含相同的文档节点，因此 D' 与 D 互为包含关系；

(6) 若 k_i 是 q 的一个查询术语，$k_i\in U$，则有一条有向边从 k_i 指向 q；

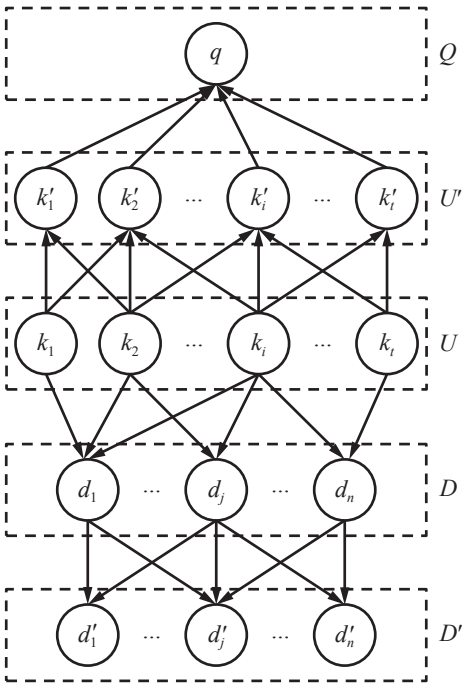


图 3 微博信念网络检索模型拓扑结构

Figure 3 Topological Structure of the Microblog Belief Network Retrieval Model

(7) 若 k_i 是 d_j 的一个索引术语，$k_i\in U$，则有一条

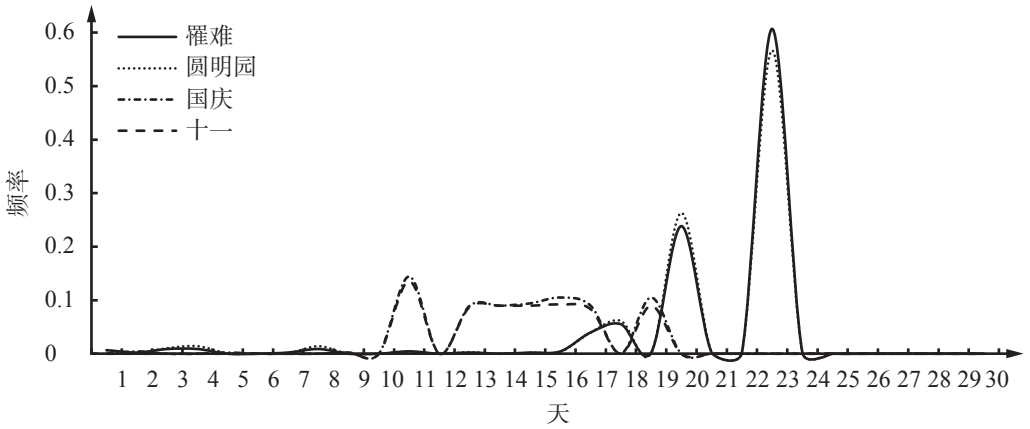


图 4 不同微博术语的词频时间分布

Figure 4 Time Distribution of Word Frequency of Different Microblog Terms

有向边从 k_i 指向 d_j;

(8) 若 k_i 与 k_i' 存在相关关系, 则有一条有向边从 k_i 指向 k_i';

(9) 若 d_j 与 d_j' 存在相关关系, 则有一条有向边从 d_j 指向 d_j'。

3.2 微博术语之间关系的量化

3.2.1 微博术语之间的相关性分析

语义信息是衡量术语相关性的主要依据[29], 如“旅行”与“旅游”两个术语具有相似语义, 它们之间存在同义关系。由于社交媒体文本具有时间敏感特性[30], 因此时间信息也是衡量微博术语之间相关性的重要依据。图 4 举例给出“罹难”“圆明园”“国庆”“十一”4 个微博术语在 30 天内的词频时间分布, 分布相似的术语之间存在较强的相关性, 如“罹难”与“圆明园”; 分布差异较大的术语之间存在较强的非相关性, 如“罹难”与“国庆”。

由此可知, 微博术语之间的相关性不仅体现在语义层面, 也体现在时间层面。为准确地量化微博术语之间的关系, 本研究混合语义与时间两方面信息计算术语之间的相关度。计算公式为

$$r_k(k_i,k_x)=\alpha r_k^{sem}(k_i,k_x)+(1-\alpha)r_k^{tim}(k_i,k_x) \tag{4}$$

其中, $r_k(k_i,k_x)$为 k_i 与 k_x 之间的相关度, k_x 为除 k_i 以外的任一术语; $r_k^{sem}(k_i,k_x)$为 k_i 与 k_x 之间的语义相关度; $r_k^{tim}(k_i,k_x)$为 k_i 与 k_x 之间的时间相关度; α为调和参数, 用于调和语义信息和时间信息对计算微博术语相关性的贡献程度。

3.2.2 微博术语之间的语义相关度

由于 Word2Vec 模型[31]能够结合语料中词语的上下文信息将一个词语映射成一个低维且稠密的语义向量, 且越相似的词语在向量空间中距离越近, 因此本研究基于 Word2Vec 模型计算微博术语之间的语义相关度。首先, 综合利用中文维基百科语料库和搜狐新闻语料库对 Word2Vec 模型进行训练; 其次, 利用训练好的 Word2Vec 模型将数据集中的微博术语转换为具有 l 维的语义向量, l 为微博术语语义向量的维度; 最后, 通过余弦距离公式计算微博术语之间的语义相关度。假定 k_i 和 k_x 为任意两个微博术语, 利用训练好的 Word2Vec 模型对其进行转换后, 可以得到 k_i 的语义向量 $\boldsymbol{s}_{k_i}$ 和 k_x 的语义向量 $\boldsymbol{s}_{k_x}$, $\boldsymbol{s}_{k_i}=(w_{i,1},\cdots,w_{i,\delta},\cdots,w_{i,l})$, $\boldsymbol{s}_{k_x}=(w_{x,1},\cdots,w_{x,\delta},\cdots,w_{x,l})$, $1\leqslant\delta\leqslant l$, $w_{i,\delta}$为 $\boldsymbol{s}_{k_i}$ 中的第 δ 维权重, $w_{x,\delta}$为 $\boldsymbol{s}_{k_x}$ 中的第 δ 维权重, 该权重即为训练好的 Word2Vec 模型的权重参数。Word2Vec 模型权重参数的训练过程[32]为: 以中文维基百科语料库和搜狐新闻语料库为训练语料, 采用连续词袋模型, 以交叉熵为目标函数, 通过梯度下降的方法对 Word2Vec 模型中的权重参数进行迭代训练, 直至收敛。采用余弦距离公式[33]计算 k_i 与 k_x 之间的语义相关度, 计算公式为

$$r_k^{sem}(k_i,k_x)=\frac{\sum_{\delta=1}^{l}w_{i,\delta}w_{x,\delta}}{\sqrt{\sum_{\delta=1}^{l}w_{i,\delta}^2}\sqrt{\sum_{\delta=1}^{l}w_{x,\delta}^2}} \tag{5}$$

3.2.3 微博术语之间的时间相关度

由于相关词汇在同一时间内呈现出相似的词频分布[34], 因此本研究依据微博术语的词频时间分布计算微博术语之间的时间相关度。首先, 以天为单位统计出每个微博术语的日出现频率, 即日出现频数与总频数的比值; 其次, 基于日出现频率将每个微博术语表示为一个时间向量; 最后, 采用余弦距离公式计算微博术语的时间相关度。对于 k_i 和 k_x, 在一定的时间范围 (o天) 内, o 为数据集的时间跨度, 同时也表示时间向量的维度数。可以得到 k_i 的时间向量 $\boldsymbol{T}_{k_i}$和 k_x 的时间向量 $\boldsymbol{T}_{k_x}$, $\boldsymbol{T}_{k_i}=(p_{i,1},\cdots,p_{i,\varepsilon},\cdots,p_{i,o})$, $\boldsymbol{T}_{k_x}=(p_{x,1},\cdots,p_{x,\varepsilon},\cdots,p_{x,o})$, $1\leqslant\varepsilon\leqslant o$, $p_{i,\varepsilon}$为 $\boldsymbol{T}_{k_i}$中的第 ε 维权重, 即 k_i 在第 ε 日的出现频率, $p_{x,\varepsilon}$为 $\boldsymbol{T}_{k_x}$中的第 ε 维权重, 即 k_x 在第 ε 日的出现频率。例如, k_i 在第 ε 日的出现频数为 10, 且其在某一时间段 o 内出现的总频数为 100, k_i 在第 ε 日出现的频率为 0.1, 即 $p_{i,\varepsilon}=0.1$。采用余弦距离公式计算 k_i 与 k_x 之间的时间相关度, 计算公

式为

$$r_k^{iim}(k_i,k_x)=\frac{\sum_{\varepsilon=1}^{o}p_{i,\varepsilon}p_{x,\varepsilon}}{\sqrt{\sum_{\varepsilon=1}^{o}p_{i,\varepsilon}^2}\sqrt{\sum_{\varepsilon=1}^{o}p_{x,\varepsilon}^2}} \tag{6}$$

3.3 微博文档之间关系量化

3.3.1 微博文档之间的相关性分析

文本的语义信息是衡量微博文档相关性的重要依据[35]，即具有相似语义的两个微博文档之间存在关系。然而，除语义信息外，微博文档还具有丰富的作者信息[36]。在本研究中作者信息即为作者发文记录。发文记录作为一种典型的作者信息，其可以有效体现微博文档的潜在主题[37]，因此作者信息在衡量微博文档相关性时是不可忽视的重要依据。图 5 给出关于 3 条微博文档及其作者词云的示例，每个词云依据相应作者的发文记录绘制，黑色箭头表示作者与微博文档的发布关系。微博文档 2 与微博文档 3 存在较强相关性，其作者词云表达的主题也比较相似；微博文档 1 与微博文档 2 和微博文档 3 之间存在较强的非相关性，作者词云表达的主题也存在明显差异。

由此可见，微博文档之间的相关性不仅体现在语义层面，而且可以通过作者信息得以体现。基于此，为准确量化微博文档关系，本研究混合语义信息和作者信息计算微博文档之间的相关度，计算公式为

$$r_d(d_j,d_y)=\beta r_d^{sem}(d_j,d_y)+(1-\beta)r_d^{aut}(d_j,d_y) \tag{7}$$

其中，$r_d(d_j,d_y)$为 d_j与 d_y之间的相关度，d_y为除 d_j以外的任一文档；$r_d^{sem}(d_j,d_y)$为 d_j与 d_y之间的语义相关度；$r_d^{aut}(d_j,d_y)$为 d_j与 d_y之间的作者相关度；β为调和参数，用于调和语义信息和作者信息对计算微博文档相关性的贡献程度。

3.3.2 微博文档之间的语义相关度

由于向量空间模型可以将文本内容转换为向量空间中的语义向量，且语义越相似的文本内容在向量空间中距离越近[27]，故本研究基于向量空间模型计算微博文档之间的语义相关度。首先，将每篇微博文档映射为一个语义向量，该语义向量由全集微博术语组成，每个维度对应一个微博术语，且每条微博文档的语义向量的维度相等，均为 t；其次，采用词频 − 逆文档频率 (term frequency-inverse document frequency, TF-IDF) 方法[38]计算语义向量中每个词语的权重；最后，采用余弦距离公式计算微博文档之间的语义相关度。假定 d_j和 d_y为任意两条微博文档，基于向量空间模型可以得到 d_j 的语义向量 $\boldsymbol{b}_{d_j}$和 d_y 的语义向量 $\boldsymbol{b}_{d_y}$，$\boldsymbol{b}_{d_j}=(c_{j,1},\cdots,c_{j,\gamma},\cdots,c_{j,t})$，$\boldsymbol{b}_{d_y}=(c_{y,1},\cdots,c_{y,\gamma},\cdots,c_{y,t})$，$1\leqslant\gamma\leqslant t$，$c_{j,\gamma}$为 $\boldsymbol{b}_{d_j}$中的第 γ 维权重，$c_{y,\gamma}$为 $\boldsymbol{b}_{d_y}$中的第 γ 维权重。以 $c_{j,\gamma}$ 的计算为例，基于 TF-IDF 方法计算权重的公式为

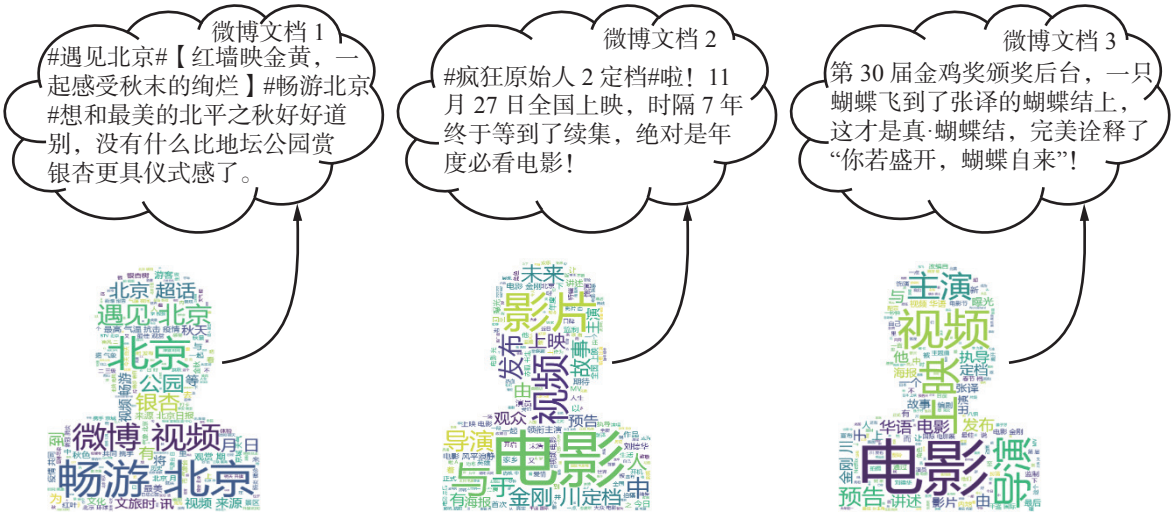
$$c_{j,\gamma}=\frac{\log\left(tf_{j,\gamma}\right)+1}{\log\left|d_j\right|}\cdot\log\frac{n}{doc_\gamma} \tag{8}$$

其中，$tf_{j,\gamma}$为 k_γ在 d_j中出现的频率，doc_γ为包含 k_γ的微博文档数，$\left|d_j\right|$为 d_j包含的词语总数。采用余弦距离公式计算 d_j与 d_y之间的语义相关度，计算公式为

$$r_d^{sem}(d_j,d_y)=\frac{\sum_{\gamma=1}^{t}c_{j,\gamma}c_{y,\gamma}}{\sqrt{\sum_{\gamma=1}^{t}c_{j,\gamma}^2}\sqrt{\sum_{\gamma=1}^{t}c_{y,\gamma}^2}} \tag{9}$$

3.3.3 微博文档之间的作者相关度

由于发文记录是表征微博用户的重要依据[39]，因



(a) 微博文档 1 的作者词云　(b) 微博文档 2 的作者词云　(c) 微博文档 3 的作者词云

图 5 微博文档及其作者词云的示例

Figure 5 An Example on Microblog Documents and Their Authors' Word Clouds

此本研究基于每个作者的微博发文记录计算作者相关度。首先，将每个作者的所有微博发文记录视作一个长文档；其次，采用向量空间模型将每个长文档表示为一个作者向量；最后，采用带有表征系数的余弦距离公式计算作者相关度。假定 h_j 为 d_j 对应的作者微博发文记录，基于向量空间模型[27]可以得到由全集微博术语组成的作者向量 $\boldsymbol{a}_{h_j}$, $\boldsymbol{a}_{h_j}=(v_{j,1},\cdots,v_{j,\theta},\cdots,v_{j,t})$, θ 为微博文档作者向量的维度, $1\leqslant\theta\leqslant t$, $v_{j,\theta}$ 为 $\boldsymbol{a}_{h_j}$ 中的第 θ 维权重；假定 h_y 为 d_y 对应的作者微博发文记录，基于向量空间模型[27]可以得到由全集微博术语组成的作者向量 $\boldsymbol{a}_{h_y}$, $\boldsymbol{a}_{h_y}=(v_{y,1},\cdots,v_{y,\theta},\cdots,v_{y,t})$, $v_{y,\theta}$ 为 $\boldsymbol{a}_{h_y}$ 中的第 θ 维权重。依据 TF-IDF 方法[27]确定权重，计算方法见 (8) 式。采用带有表征系数的余弦距离公式计算 d_j 与 d_y 之间的作者相关度，计算公式为

$$r_d^{aut}(d_j,d_y)=\mu_j\mu_y\frac{\sum_{\theta=1}^{t}v_{j,\theta}v_{y,\theta}}{\sqrt{\sum_{\theta=1}^{t}v_{j,\theta}^2}\sqrt{\sum_{\theta=1}^{t}v_{y,\theta}^2}} \tag{10}$$

其中，μ_j 为 $\boldsymbol{a}_{h_j}$ 对 d_j 的表征能力系数，μ_y 为 $\boldsymbol{a}_{h_y}$ 对 d_y 的表征能力系数，该系数用以控制作者发文活跃性[40]对作者相关度计算的影响。本研究令 $\mu_j=\frac{1}{|h_j|}$、$\mu_y=\frac{1}{|h_y|}$, 其具体含义为：当作者的微博发文记录越多时，作者向量对某一微博文档的表征能力越弱。本研究引入表征能力系数，旨在更准确地利用作者向量对微博文档之间的关系进行量化。

3.4 推理与检索

依据微博信念网络检索模型的拓扑结构，模型的检索过程即计算 $P(d'_j|q)$ 的过程。依据基本信念网络检索给出的检索过程，即 (1) 式，本研究将计算 $P(d'_j|q)$ 的过程表示为

$$P(d'_j|q)=\eta\sum_{\forall u}P(d'_j|u)P(q|u)P(u) \tag{11}$$

由于 u 与 q 之间有一层术语节点 k'_i，需要重新定义 $P(q|u)$ 的计算。依据贝叶斯概率推导[7]，计算 $P(q|u)$ 的公式为

$$P(q|u)=\sum_{\forall u'}P(q|u')P(u'|u) \tag{12}$$

其中，概念 u' 由概念 u 复制而来，为概念空间 U' 的一个子集。

由于 u 与 d'_j 之间有一层文档节点 d_j，需要重新定义 $P(d'_j|u)$ 的计算。依据贝叶斯概率推导，计算 $P(d'_j|u)$ 的公式为

$$P(d'_j|u)=\sum_{d_j\in fa(d'_j)}P(d_j|u)P(d'_j|d_j) \tag{13}$$

其中，$fa(d'_j)$ 为 d'_j 的父节点集合。

把 (12) 式和 (13) 式代入 (11) 式，得到

$$\begin{aligned}P(d'_j|q)&=\eta\sum_{\forall u}[\sum_{d_j\in fa(d'_j)}P(d_j|u)P(d'_j|d_j)]\cdot\\&\quad[\sum_{\forall u'}P(q|u')P(u'|u)]P(u)\\&=\eta\sum_{\forall u}\sum_{\forall u'}P(q|u')P(u'|u)\cdot\\&\quad[\sum_{d_j\in fa(d'_j)}P(d_j|u)P(d'_j|d_j)]P(u)\end{aligned} \tag{14}$$

其中，$P(q|u')$ 和 $P(d_j|u)$ 可依据 (1) 式的基本信念网络检索模型进行计算。对于 $P(u'|u)$ 和 $P(d'_j|d_j)$ 的计算，为有效利用量化的微博文档关系，令 $P(d'_j|d_j)=r(d'_j,d_j)$, 且 $r(d'_j,d_j)$ 依据 (7) 式确定。为有效利用微博术语关系，计算 $P(u'|u)$ 的公式为

$$P(u'|u)=\frac{1}{m}\sum_{\forall k'_i\in u'\,fa(k'_i)\in u}P[k'_i|fa(k'_i)] \tag{15}$$

其中，m 为 u' 中包含的术语个数。可以将 (15) 式理解为：u' 包含的每一个 k'_i 在其父节点条件下的概率平均值。进一步，计算 $P[k'_i|fa(k'_i)]$ 的公式为

$$P[k'_i|fa(k'_i)]=\frac{\sum_{k_j\in fa(k'_i)}r(k'_i,k_j)}{|fa(k'_i)|} \tag{16}$$

其中，$|fa(k'_i)|$ 为 k'_i 父节点的个数，$r(k'_i,k_j)$ 可依据 (4) 式确定。

4 实验

4.1 实验数据

目前，微博检索中比较权威的实验数据集为文本检索会议 (Text REtrieval Conference, TREC) 的评测数据集[41−42]，但此数据集包含的字段信息有限，无法满足本研究的实验要求。为对本研究提出的模型进行有效验证，依据中小型信息检索测试集的构建标准[43]整理出一个符合实验需求的数据集。本研究采用网络爬虫工具对真实的新浪微博数据进行爬取：①选定 5 个真实的微博话题数据集，并分别爬取每个话题从 2020 年 10 月 7 日至 2020 年 11 月 7 日的微博信息，包括发布者 ID、博文 ID、发布时间和微博内容 4 个字段；②整理出每个话题中包含的所有发布者 ID，并分别爬取每个发布者最近发布的 300 条历史微博，每条历史微博包含发布者 ID、博文 ID、发布时间和微博内容 4 个字段的信息。基于爬取的原始微博数据，参照 TREC 的评测标准，对原始微博数据进行预处理：①去除已失效或不存在的微博；②去除长度小于 40 个字符的微博；③将微博中繁体中文转化为简体中文。通过上述操作，获得表 1 所统计的数据信息。其中，将每个话题视为查询语句，将每个话题下出现的微博文档视作对应查询的相关文档，将每个话题下未出现的微博文档视作对应查询的不相关文档。

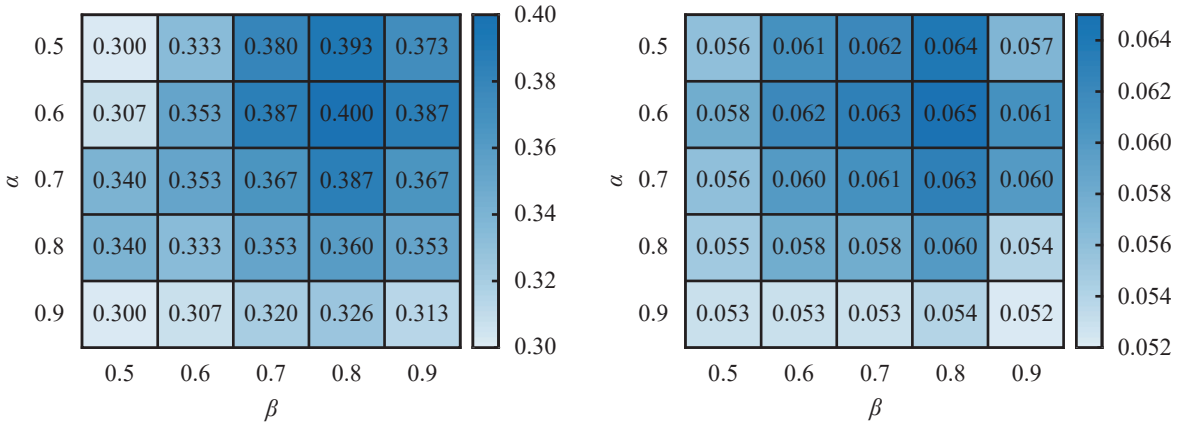
4.2 评价指标

参照文本检索会议的评测标准[42]和已有研究[44]，本研究采用前 K 个返回结果的准确率和平均查准率对微博检索性能进行评价，准确率强调查询结果的

表1 微博数据集统计

Table 1 Statistics of Microblog Datasets

编号	话题词	相关文档数	不相关文档数	涉及作者数	作者发文记录
1	故宫600年	4 059	288 293	250	21 153
2	国庆节	27 017	265 335	245	17 045
3	美国疫情	17 550	274 802	249	35 792
4	七夕	9 694	282 658	244	19 507
5	电影花木兰	958	291 394	244	23 260



(a) 不同参数对应的 *Pre*@30 (b) 不同参数对应的 *Map*@30

图6 不同参数组合对应的检索性能

Figure 6 Retrieval Performance Corresponding to Different Parameter Combinations

准确率，平均查准率在准确率的基础上考虑了位置因素，可有效衡量相关文档的排名顺序。具体计算公式为

$$Pre@K = \frac{1}{K}\sum_{z=1}^{K} rel_z \tag{17}$$

$$Map@K = \frac{1}{R}\left(\frac{1}{K}\sum_{z=1}^{K}\frac{1}{Ran_z}\right) \tag{18}$$

其中，z 为查询结果，$z = 1,\cdots,K$；$Pre@K$ 为前 K 个返回的查询结果的准确率；$Map@K$ 为前 K 个返回的查询结果的平均查准率；R 为查询主题的数量；rel_z 为查询结果的相关性，即当第 z 个查询结果与查询需求相关时 $rel_z = 1$，否则 $rel_z = 0$；Ran_z 为返回结果中第 z 个查询结果的位置。由于微博检索首页通常包含15条结果，且微博用户通常仅关注前两页的检索结果[40]，因此本研究分别令 K 取值为15和30，对不同模型的性能进行比较，$Pre@K$ 和 $Map@K$ 的值越高，表示检索性能越好。

4.3 实验和分析

本研究实验包括3个部分，第1部分确定相关参数，第2部分对比本研究提出的模型与主流模型的性能，第3部分对本研究提出的模型进行消融实验，即验证不同部分对模型性能的影响。

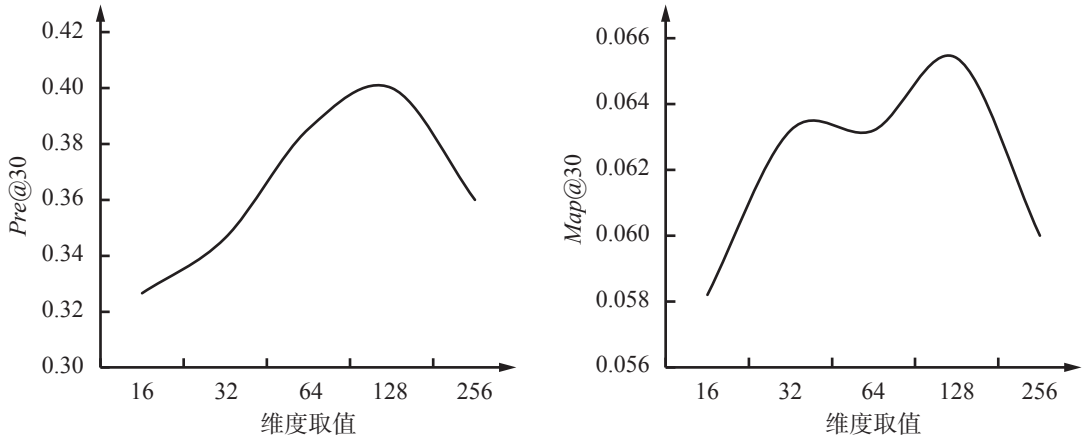
4.3.1 确定参数

(1) 确定调和参数 α 和 β

本研究通过网格搜索的方法确定 α 和 β 的较优取值。由于语义信息是相关性认知的关键且稳定的依据，因此本研究分别令 α、β 在 {0.5, 0.6, 0.7, 0.8, 0.9} 中取值，并依据微博信念网络检索模型的性能确定较优取值。图6给出 α 和 β 取不同值时对应的检索性能，当 α 取值为0.6且 β 取值为0.8，α 和 β 分别取值为0.7和0.8时，微博信念网络建模在 *Pre*@30 和 *Map*@30 上均体现出较优的检索性能。由于 *Pre*@30 和 *Map*@30 的值越大，模型的检索性能越好。因此，本研究将 α 取值为0.6，β 取值为0.8。

(2) 确定维度参数 l

维度参数 l 即基于Word2Vec模型的术语语义向量维度，其关系微博术语语义表示的准确性。由于图形处理单元通常在16的倍数时发挥较优性能，因此本研究分别令 l 在 {16, 32, 64, 128, 256} 中取值，并依据微博信念网络检索模型的性能确定参数的较优取值。图7给出 l 取不同值时对应的微博信念网络检索模型的性能，当 l 取值128时，微博信念网络检索模型在 *Pre*@30 和 *Map*@30 上均体现出较优的检索



(a) 不同维度对应的 *Pre*@30　　(b) 不同维度对应的 *Map*@30

图 7 不同维度参数对应的检索性能

Figure 7 Retrieval Performance Corresponding to Parameters of Different Dimensions

表 2 不同模型的性能比较

Table 2 Performance Comparison of Different Models

	AQLM	MQLM	UHMR	MBNR	*Imp*/%
Pre@15	0.280	0.347	0.387	0.467	20.672
Pre@30	0.267	0.353	0.340	0.400	13.314
Map@15	0.052	0.062	0.091	0.115	26.374
Map@30	0.032	0.040	0.052	0.065	25

注: *Imp*为MBNR模型相对于次优模型的性能提升程度, 下同。

性能。因此, 本研究将术语的维度参数 l 确定为 128。

4.3.2 性能比较

为验证本研究构建的模型的有效性和合理性, 选取 3 个主流的微博检索模型作为基线模型, 与本研究构建模型的性能进行对比实验。3 个基线模型和本研究模型的解释如下:

(1) 基于作者建模的微博检索模型 (AQLM)[39]。该模型将每个作者发布的所有微博记录视作原始微博平滑项, 并基于语言模型重新对微博文档中的词语概率进行估计, 以结合作者信息缓解微博文本固有的特征稀疏性。

(2) 多源信息融合的微博查询似然模型 (MQLM)[11]。该模型在查询似然模型的基础上, 综合微博自身、微博主题和微博全集等多源信息对微博语言模型估计进行改进, 不仅避免了语言模型估计面临的零概率问题, 而且通过差异化处理语言模型的平滑项, 提升了微博文档中词项估计的准确性。

(3) 融合用户兴趣和混合估计的微博检索模型 (UHMR) [25]。该模型在查询似然模型的基础上, 一方面结合用户兴趣改进了微博文档的先验概率, 另一方面结合用户交互改进了微博语言模型的估计, 不仅满足了用户的个性化检索需求, 而且有效扩展了微博文档的语义特征。

(4) 多粒度融合的微博信念网络检索模型 (MBNR)。该模型即为本研究提出的微博检索模型, 一方面混合时间信息和语义信息对微博术语之间关系进行量化, 另一方面混合作者信息和语义信息对微博文档之间关系进行量化, 结合量化后的术语关系和文档关系在信念网络模型的基础上实现微博检索。

表 2 给出基线模型与本研究提出模型的检索性能比较结果, 与 AQLM、MQLM 和 UHMR 相比, MBNR 的 *Pre*@15、*Pre*@30、*Map*@15 和 *Map*@30 的值最高, 表明 MBNR 具有较好的微博检索性能。产生这样结果的原因是: ①AQLM 和 MQLM 仅对微博文档中词项概率的估计进行改进, 虽然有效缓解了微博文档固有的语义稀疏性问题, 但忽视了较短的微博查询语句对微博检索性能造成的消极影响。与 AQLM 和 MQLM 相比, 一方面 MBNR 利用术语之间关系有效扩展了较短的微博查询语句, 另一方面利用文档之间关系有效扩展了微博文档, 因此 MBNR 可以同时缓解较短的查询语句和文档对检索性能造成的消极影响。②UHMR 虽然通过改进微博文档的先验概率有效考虑了用户的个性检索需求, 但在本质上仍是基于用户提供的微博查询语句实现微博文

档匹配，并未对微博查询语句进行有效扩展。与UHMR相比，MBNR同时利用微博术语的语义和时间两方面的相关性对原始的微博查询语句进行扩展，可以更充分地考虑用户的个性化检索需求。

4.3.3 消融分析

为验证在信念网络检索模型的基础上同时考虑术语之间关系和文档之间关系的有效性，本研究通过逐步删除关键部分将MBNR模型演变为3个不同模型，并通过消融实验验证研究工作的有效性。由MBNR演变的3个模型的解释如下：

(1) 基本信念网络检索模型(BNR)。该模型仅用基本信念网络检索模型实现微博检索任务，与MBNR相比，BNR既未考虑微博术语之间的关系，也未考虑微博文档之间的关系，仅通过查询语句与文档的关键词匹配实现微博检索。

(2) 考虑术语关系的信念网络检索模型(TBNR)。该模型在基本信念网络检索模型的基础上融合微博术语之间的关系实现微博检索任务，与MBNR相比，TBNR未考虑微博文档之间的关系。此外，为保证模型之间的可比性，对TBNR中微博术语之间关系的量化方法与MBNR的一致。

(3) 考虑文档关系的信念网络检索模型(DBNR)。该模型在基本信念网络检索模型的基础上融合微博文档之间的关系实现微博检索任务，与MBNR相比，DBNR未考虑微博术语之间的关系。此外，为保证模型之间的可比性，对DBNR中微博文档之间关系的量化方法与MBNR的一致。

表3给出消融模型与本研究提出模型的检索性能比较结果，与BNR、TBNR和DBNR相比，MBNR的*Pre*@15、*Pre*@30、*Map*@15和*Map*@30的值最高，表明在信念网络检索模型的基础上，同时融合微博术语之间的关系和文档之间的关系可以有效提升微博检索性能。产生这样结果的原因是：①BNR仅能依据关键词匹配实现微博检索功能，不能避免较短的微博查询语句和微博文档对检索性能造成的消极影响；②TBNR仅利用微博术语之间的关系实现了微博查询扩展的目的，忽略了微博文档之间的关系对微博检索性能的辅助提升作用；③DBNR仅利用微博文档之间的关系实现了微博文档扩展的目标，忽略了微博术语之间的关系对微博检索性能的辅助提升作用。与BNR、TBNR和DBNR相比，一方面MBNR利用术语之间的关系有效扩展了原始微博查询语句，另一方面利用文档之间的关系有效扩展了原始微博文档，因此MBNR具有较好的微博检索性能。

表3 消融实验结果

Table 3 Results of Ablation Experiments

	BNR	TBNR	DBNR	MBNR	*Imp*/%
Pre@15	0.187	0.387	0.320	0.467	20.672
Pre@30	0.173	0.307	0.333	0.400	20.120
Map@15	0.027	0.073	0.053	0.115	57.534
Map@30	0.017	0.042	0.034	0.065	54.762

5 结论

5.1 研究结果

微博信息过载的背景下，设计合理的微博检索系统是提升微博信息服务水平的关键。针对微博检索中查询语句和文档长度较短导致的词不匹配问题，本研究结合微博术语和文档的特点，在信念网络检索模型的基础上，通过融合术语关系和文档关系，提出一个多粒度融合的微博信念网络检索模型，通过网络爬虫工具获取真实的新浪微博数据，对模型的合理性和有效性进行实证检验。研究结果表明，①与AQLM模型、MQLM模型、UHMR模型等当前主流基线模型中表现较好的指标进行对比，本研究提出的MBNR模型在*Pre*@15、*Pre*@30、*Map*@15和*Map*@30等4项指标上分别提升了20.672%、13.314%、26.374%和25%，表明MBNR模型可以表现出较优的微博检索性能。②为检验综合利用术语关系和文档关系扩展信念网络检索模型的有效性，根据MBNR模型衍生出BNR模型、TBNR模型和DBNR模型进行消融实验，结果表明与表现较好的TBNR模型和DBNR模型相比，MBNR模型在*Pre*@15、*Pre*@30、*Map*@15和*Map*@30等4项指标上分别提升了20.672%、20.120%、57.534%和54.762%。显著的性能提升表明，在信念网络检索模型的基础上，同时融合术语关系和文档关系可以有效提升信念网络检索模型在微博检索情景下的适用性。

5.2 研究价值

本研究的理论价值主要体现为：①结合微博检索面临的挑战，在信念网络检索模型的基础上，综合术语之间的关系和文档之间的关系，提出一个多粒度融合的微博信念网络检索模型，不仅拓展了信念网络检索模型在微博检索研究中的适用性，而且结合查询语句扩展和文档扩展的优势有效提高了微博检索的性能。②结合微博术语和微博文档的特点，一方面提出一种微博术语之间关系的量化方法，另一方面提出一种微博文档之间关系的量化方法，不仅提高了测量微博术语之间的关系和文档之间的关系的准确性，而且有利于使本研究提出的微博信念网络检索模型更好地发挥微博检索性能。本研究的实践价值主要体现为：聚焦于微博平台面临的信息过载难题，通过设计微博检索模型有效地实现了信息过滤，辅助用户从海量的微博中获取个性化且感兴趣的高质量信息，这不仅有助于为微博平台开发合理的信息检索系统提供思路，而且为社交媒体平台增强用户使用体验、提升信息服务水平提供了技术借鉴。

5.3 研究局限和展望

与已有研究相比，本研究虽在一定程度上提高了微博检索的性能，但尚存不足之处。①本研究在量

化微博术语关系时仅考虑了语义和时间两方面信息，在量化微博文档关系时仅考虑语义和作者两方面信息。然而，微博术语之间或微博文档之间的相关性还可能与其他多种上下文信息有关，如微博术语的语法相关性、微博文档的话题相关性或时空相关性。因此，未来研究可以尝试分析影响术语相关性和文档相关性的多方面因素，并结合多源信息对微博术语之间的关系和文档之间的关系进行更准确的量化。②本研究针对的微博检索对象为微博文档，因此仅强调了文本模态的信息。然而，随着移动互联网技术的快速发展，微博已成为一种综合文本、音频和影像等多模态信息的综合体，不同模态的信息有助于从不同角度对微博的表示进行建模。因此，未来研究可以尝试综合多模态信息对微博进行建模，以提高微博检索系统的多模态感知能力。③本研究在实现微博检索的过程中未能有效对微博网络术语的语义信息进行翻译，如“YYDS”表示“永远的神”、“针不戳”表示“真不错”等，因此当微博中包含大量的网络术语时无法较好地实现微博语义检索。未来研究可以对微博网络术语进行识别，并通过对其语义翻译实现更有效的微博语义检索。

参考文献：

[1] GARCIA K, BERTON L. Topic detection and sentiment analysis in Twitter content related to COVID-19 from Brazil and the USA. ***Applied Soft Computing***, 2021, 101: 107057-1−107057-15.

[2] GAO J M, ZHANG C X, XU Y Y, et al. Hybrid microblog recommendation with heterogeneous features using deep neural network. ***Expert Systems with Applications***, 2021, 167: 114191-1−114191-13.

[3] HAN Z Y, KONG L L, QI H L. Time segment language model for microblog retrieval. ***Neural Computing and Applications***, 2021, 33(10): 4763−4777.

[4] ZINGLA M A, LATIRI C, MULHEM P, et al. Hybrid query expansion model for text and microblog information retrieval. ***Information Retrieval Journal***, 2018, 21(4): 337−367.

[5] GAN M X, ZHANG X T. Integrating community interest and neighbor semantic for microblog recommendation. ***International Journal of Web Services Research***, 2021, 18(2): 54−75.

[6] 黄名选. 关联模式挖掘与词向量学习融合的伪相关反馈查询扩展. **电子学报**, 2021, 49(7): 1305−1313.
HUANG Mingxuan. Pseudo-relevance feedback query expansion based on the fusion of association pattern mining and word embedding learning. ***Acta Electronica Sinica***, 2021, 49(7): 1305−1313.

[7] DE CRISTO M A P, CALADO P P, DE LOURDES DA SILVEIRA M, et al. Bayesian belief networks for IR. ***International Journal of Approximate Reasoning***, 2003, 34(2/3): 163−179.

[8] 徐建民, 何丹丹, 吴树芳. 基于文档关系的扩展信念网络检索模型. **情报学报**, 2019, 38(11): 1160−1165.
XU Jianmin, HE Dandan, WU Shufang. An extended belief network retrieval model based on document relationships. ***Journal of the China Society for Scientific and Technical Information***, 2019, 38(11): 1160−1165.

[9] ZHU X, HUANG J M, ZHOU B, et al. Real-time personalized Twitter search based on semantic expansion and quality model. ***Neurocomputing***, 2017, 254: 13−21.

[10] NASIR J A, VARLAMIS I, ISHFAQ S. A knowledge-based semantic framework for query expansion. ***Information Processing & Management***, 2019, 56(5): 1605−1617.

[11] 吴树芳, 张雄涛, 朱杰. 多源信息融合的微博查询似然模型. **图书情报工作**, 2020, 64(17): 114−122.
WU Shufang, ZHANG Xiongtao, ZHU Jie. Microblog query likelihood model based on multi-source information fusion. ***Library and Information Service***, 2020, 64(17): 114−122.

[12] 韩中元, 杨沐昀, 孔蕾蕾, 等. 基于词汇时间分布的微博查询扩展. **计算机学报**, 2016, 39(10): 2031−2044.
HAN Zhongyuan, YANG Muyun, KONG Leilei, et al. Query expansion based on term time distribution for microblog retrieval. ***Chinese Journal of Computers***, 2016, 39(10): 2031−2044.

[13] 安璐, 胡俊阳, 李纲. 基于主题一致性和情感支持的评论意见领袖识别方法研究. **管理科学**, 2019, 32(1): 3−13.
AN Lu, HU Junyang, LI Gang. A method of identifying comment opinion leaders based on topic consistency and emotional support. ***Journal of Management Science***, 2019, 32(1): 3−13.

[14] 严建援, 李扬, 冯淼, 等. 用户问答与在线评论对消费者产品态度的交互影响. **管理科学**, 2020, 33(2): 102−113.
YAN Jianyuan, LI Yang, FENG Miao, et al. Interaction effects of customer Q&As and online reviews on consumer product attitudes. ***Journal of Management Science***, 2020, 33(2): 102−113.

[15] XU F, SHENG V S, WANG M W. Near real-time topic-driven rumor detection in source microblogs. ***Knowledge-Based Systems***, 2020, 207: 106391-1−106391-9.

[16] 刘蕾, 于春玲, 赵平. 图文信息对消费者互动行为及品牌关系的影响. **管理科学**, 2018, 31(1): 90−100.
LIU Lei, YU Chunling, ZHAO Ping. Impact of picture-word information on consumer engagement behavior and consumer-brand relationship. ***Journal of Management Science***, 2018, 31(1): 90−100.

[17] ABDALGADER K, AL SHIBLI A. Context expansion approach for graph-based word sense disambiguation. ***Expert Systems with Applications***, 2021, 168: 114313-1−114313-15.

[18] KALLOUBI F, NFAOUI E H, EL BEQQALI O. Microblog semantic context retrieval system based on linked open data and graph-based theory. ***Expert Systems with Applications***, 2016, 53: 138−148.

[19] SAMUEL A, SHARMA D K. A spatial, temporal and sentiment based framework for indexing and clustering in Twitter blogosphere. ***Journal of Intelligent & Fuzzy Systems***, 2017, 32(5): 3619−3632.

[20] 寇菲菲, 杜军平, 石岩松, 等. 面向搜索的微博短文本语义建模方法. **计算机学报**, 2020, 43(5): 781−795.
KOU Feifei, DU Junping, SHI Yansong, et al. Microblog short text semantic modeling method for search. ***Chinese Journal of Computers***, 2020, 43(5): 781−795.

[21] 熊才伟, 曹亚男. 基于发文内容的微博用户兴趣挖掘方法研究. **计算机应用研究**, 2018, 35(6): 1619−1623.
XIONG Caiwei, CAO Yanan. Research of microblog user interest mining based on microblog posts. ***Application Research of Computers***, 2018, 35(6): 1619−1623.

[22] SPINA D, ZUBIAGA A, SHETH A, et al. Processing social media

in real-time. ***Information Processing & Management***, 2019, 56(3): 1081−1083.

[23] ZHOU N, DU J P, YAO X, et al. A content search method for security topics in microblog based on deep reinforcement learning. ***World Wide Web***, 2020, 23(1): 75−101.

[24] TIAN Y, ZHOU L J, ZHANG Y, et al. Deep cross-modal face naming for people news retrieval. ***IEEE Transactions on Knowledge and Data Engineering***, 2021, 33(5): 1891−1905.

[25] 吴树芳, 张雄涛, 朱杰. 融合用户兴趣和混合估计的微博检索模型. **情报学报**, 2019, 38(4): 411−419.
WU Shufang, ZHANG Xiongtao, ZHU Jie. Microblog retrieval model combining user interest and mixed estimation. ***Journal of the China Society for Scientific and Technical Information***, 2019, 38(4): 411−419.

[26] 郑伟, 侯宏旭, 武静. 贝叶斯网络在信息检索中的应用. **情报科学**, 2018, 36(6): 136−141.
ZHENG Wei, HOU Hongxu, WU Jing. Application of Bayesian network for information retrieval. ***Information Science***, 2018, 36(6): 136−141.

[27] BELWAL R C, RAI S, GUPTA A. Text summarization using topic-based vector space model and semantic measure. ***Information Processing & Management***, 2021, 58(3): 102536-1−102536-15.

[28] 白彦霞, 程杰, 莫德举. 基于语义扩展的复合贝叶斯网络检索模型. **控制工程**, 2019, 26(3): 602−607.
BAI Yanxia, CHENG Jie, MO Deju. Compound Bayesian network retrieval model based on semantic extension. ***Control Engineering of China***, 2019, 26(3): 602−607.

[29] 刘阳光, 岂凡超, 刘知远, 等. HowNet义原标注一致性检验方法研究. **中文信息学报**, 2021, 35(4): 23−34.
LIU Yangguang, QI Fanchao, LIU Zhiyuan, et al. Research on consistency check of sememe annotations in HowNet. ***Journal of Chinese Information Processing***, 2021, 35(4): 23−34.

[30] 石磊, 陶永才, 李俊艳, 等. 个性化微博实时推荐模型研究. **小型微型计算机系统**, 2016, 37(9): 1910−1914.
SHI Lei, TAO Yongcai, LI Junyan, et al. Personalized and real-time recommendation model for microblogs. ***Journal of Chinese Computer Systems***, 2016, 37(9): 1910−1914.

[31] SHIVAKUMAR P G, GEORGIOU P. Confusion2Vec: towards enriching vector space word representations with representational ambiguities. ***Peer J Computer Science***, 2019, 5: e195-1−e195-49.

[32] KHATUA A, KHATUA A, CAMBRIA E. A tale of two epidemics: contextual word2Vec for classifying Twitter streams during outbreaks. ***Information Processing & Management***, 2019, 56(1): 247−257.

[33] XIA P P, ZHANG L, LI F Z. Learning similarity with cosine similarity ensemble. ***Information Sciences***, 2015, 307: 39−52.

[34] LI X, JIANG M M, HONG H T, et al. A time-aware personalized point-of-interest recommendation via high-order tensor factorization. ***ACM Transactions on Information Systems***, 2017, 35(4): 1−23.

[35] 朱茂然, 朱艳鹏, 高松, 等. 基于深度哈希的相似图片推荐系统：以Airbnb为例. **管理科学**, 2020, 33(5): 17−28.
ZHU Maoran, ZHU Yanpeng, GAO Song, et al. Similar picture recommendation system based on deep hashing: evidence from the Airbnb platform. ***Journal of Management Science***, 2020, 33(5): 17−28.

[36] XU D L, TIAN Z H, LAI R F, et al. Deep learning based emotion analysis of microblog texts. ***Information Fusion***, 2020, 64: 1−11.

[37] 闫志华, 唐锡晋. 融合高效用模式的在线媒体突发话题发现. **系统工程理论与实践**, 2021, 41(5): 1138−1149.
YAN Zhihua, TANG Xijin. Bursty topic discovery of online media incorporating high utility pattern. ***Systems Engineering - Theory & Practice***, 2021, 41(5): 1138−1149.

[38] JIANG Z Y, GAO B, HE Y L, et al. Text classification using novel term weighting scheme-based improved TF-IDF for internet media reports. ***Mathematical Problems in Engineering***, 2021: 6619088-1−6619088-30(Online).

[39] 李锐, 王斌. 一种基于作者建模的微博检索模型. **中文信息学报**, 2014, 28(2): 136−143.
LI Rui, WANG Bin. Microblog retrieval via author based microblog expansion. ***Journal of Chinese Information Processing***, 2014, 28(2): 136−143.

[40] 周静, 沈俏蔚, 涂平, 等. 社交网络中用户关注类型与发帖类型对发帖行为的影响. **管理科学**, 2019, 32(2): 67−76.
ZHOU Jing, SHEN Qiaowei, TU Ping, et al. Influence of neighbor and posting type on users′ posting behavior in social networks. ***Journal of Management Science***, 2019, 32(2): 67−76.

[41] AKLOUCHE B, BOUNHAS I, SLIMANI Y. BM25 beyond query-document similarity//***Proceedings of the 26th International Symposium on String Processing and Information Retrieval***. Segovia: Springer, 2019: 65−79.

[42] LIN Y, XU B, LIN H F, et al. FGFIREM: a feature generation framework based on information retrieval evaluation measures. ***Expert Systems with Applications***, 2019, 133: 75−85.

[43] 徐建民, 王平. 小型中文信息检索测试集的构建与分析. **情报杂志**, 2009, 28(1): 13−16.
XU Jianmin, WANG Ping. Small Chinese information retrieval test collections: construction and analysis. ***Journal of Intelligence***, 2009, 28(1): 13−16.

[44] ALBISHRE K, LI Y F, XU Y, et al. Query-based unsupervised learning for improving social media search. ***World Wide Web***, 2020, 23(3): 1791−1809.

A Microblog Belief Network Retrieval Model Integrating Multi-granularity Relationships

ZHANG Xiongtao, GAN Mingxin, LI Shuo
School of Economics and Management, University of Science and Technology Beijing, Beijing 100083, China

Abstract: The microblog retrieval system is an important tool for microblog platform to realize personalized information filtering. Establishing a reasonable microblog retrieval model is not only conducive to meet users′ personalized information demands, but also to improve the information service level of microblog platform. However, compared with the traditional text retrieval, microblog retrieval faces two challenges: on the one hand, the shorter query is difficult to accurately express the user′s retrieval demands, yet on the other hand, the shorter microblog is difficult to fully express semantic, which makes difficult to accurately match queries and documents.

Combining the characteristics of microblog terms and microblog documents, the relationships between terms and documents are integrated into the belief network retrieval model, and a microblog belief network retrieval model integrating multi-granularity relationships is proposed. Firstly, the relationship between microblog terms is quantified by mixing semantic information and temporal information, so as to model the relevance between microblog terms more accurately. Secondly, the relationship between microblog documents is quantified by mixing semantic information and author information, so as to model the relevance between microblog documents more accurately. Finally, based on the basic belief network retrieval model, the probabilistic derivation process of microblog belief network retrieval model is given by combining the quantitative term relationship and document relationship. The study uses web crawler to obtain real microblog data from Sina Weibo to verify the validity and rationality of microblog belief network retrieval model.

The results show that there are semantic relevance and temporal relevance between microblog terms, while there are semantic relevance and author relevance between microblog documents. In addition, the results also show that compared with the mainstream microblog retrieval models, the microblog belief network retrieval model has better retrieval performance in multiple information retrieval metrics. Furthermore, the ablation experiment results show that the belief network retrieval model integrating multi-granularity relationship has better retrieval performance than the model considering single granularity relationship or no relationship at all.

The study focuses on the microblog retrieval scenario and integrated the advantages of query extension and document extension to achieve microblog retrieval, which effectively overcomes the challenges faced by microblog retrieval and significantly improves the performance of microblog retrieval. In the context of information overload, the microblog belief network retrieval model not only helps to further improve the information service level of social media platforms, but also provides reference for the development of a reasonable retrieval system for social media platforms.

Keywords: multi-granularity relationship integration; microblog retrieval; belief network retrieval model; term relationship; document relationship

Received Date: August 27th, 2021　　**Accepted Date:** May 10th, 2022

Funded Project: Supported by the National Natural Science Foundation of China (72271024, 71871019)

Biography: ZHANG Xiongtao is a Ph.D candidate in the School of Economics and Management at University of Science and Technology Beijing. His research interests include social media computing, information retrieval and recommendation. His representative paper titled "Integrating community interest and neighbor semantic for microblog recommendation" was published in the ***International Journal of Web Services Research*** (Issue 2, 2021). E-mail: B20190412@xs.ustb. edu.cn

GAN Mingxin, doctor in management, is a professor in the School of Economics and Management at University of Science and Technology Beijing. Her research interests include social media computing and recommendation systems. Her representative paper titled "A knowledge-enhanced contextual bandit approach for personalized recommendation in dynamic domain" was published in the ***Knowledge-Based Systems*** (Volume 251, 2022). E-mail: ganmx@ustb.edu.cn

LI Shuo is a master degree candidate in the School of Economics and Management at University of Science and Technology Beijing. His research interests include social media computing and intelligent data analysis. E-mail: 18511827966@163.com

(责任编辑: 李祎博)