中国科学: 信息科学 2020 年 第 50 卷 第 6 期: 845–861
SCIENTIA SINICA Informationis
大数据分析专题 · 论文

一种高可读低冗余实体摘要的生成方法

刘庆霞, 程龚*, 瞿裕忠

南京大学计算机软件新技术国家重点实验室, 南京 210023
* 通信作者. E-mail: gcheng@nju.edu.cn

收稿日期: 2019–12–30; 修回日期: 2020–03–08; 接受日期: 2020–04–09; 网络出版日期: 2020–06–08

国家重点研发计划 (批准号: 2018YFB1004300) 和国家自然科学基金 (批准号: 61772264) 资助项目

摘要 随着万维网的发展, 知识图谱数据大量增长, 并在面向智能应用的研究中受到广泛关注. 知识图谱用 RDF (resource description framework) 三元组描述实体相关的事实. 在知识图谱中, 关于一个实体的描述可能包含大量三元组, 在一些需要直接呈现实体信息的应用中, 为了避免用户信息过载, 并适应有限的呈现空间, 就需要进行实体摘要. 实体摘要任务是从实体描述的众多三元组中选出最有代表性的子集作为摘要, 以呈现给用户阅读. 本文提出一种新的实体摘要方法 ESSTER 以生成具备高可读性和低冗余性的实体摘要. 该方法结合三元组的结构与文本特征, 基于结构特性度量知识图谱中三元组的重要性, 基于 N 元语法和文本语料度量三元组的可读性, 基于逻辑推理、数值比较和文本相似判断三元组间的冗余关系. 综合这 3 种技术要素, 将实体摘要问题建模为组合优化问题进行求解. 本文在两个由人工标注的公开数据集上与 6 种现有方法进行了对比实验, 结果表明本文提出的方法效果达到了当前最佳水平.

关键词 知识图谱, 实体摘要, 冗余性, 可读性, 组合优化

1 引言

利用知识图谱中的丰富知识满足用户的信息需求, 是近年来智能应用研究的一个重要话题. 众多研究机构和企业先后构建了大量知识图谱, 例如 Google 的 Knowledge Graph, Facebook 的 Open Graph, 微软的 Concept Graph, 以及 DBpedia, Freebase, LinkedMDB 等. 知识图谱以资源描述框架 (resource description framework, RDF) 数据模型来表达知识, 该模型可将知识呈现为图结构. 图的节点可以表示各种类型的实体 (例如人物、地点等) 或字面量 (如文本、日期、数值等), 图中的有向边表示实体与实体之间, 或者实体与字面量之间存在的各种联系 (例如出生地、年龄等). 知识图谱中由节点和有向边构成的三元组 ⟨ 主语, 谓词, 宾语 ⟩ 描述了现实中的一条事实. 实体可以出现在主语或宾语位置. 本

引用格式: 刘庆霞, 程龚, 瞿裕忠. 一种高可读低冗余实体摘要的生成方法. 中国科学: 信息科学, 2020, 50: 845–861, doi: 10.1360/SSI-2019-0291
Liu Q X, Cheng G, Qu Y Z. Entity summarization with high readability and low redundancy (in Chinese). Sci Sin Inform, 2020, 50: 845–861, doi: 10.1360/SSI-2019-0291

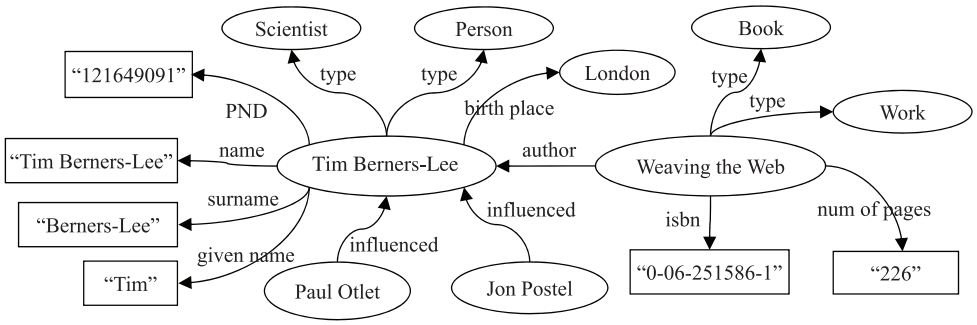

图 1　知识图谱示例 (实体和类型用椭圆表示, 字面量用矩形框表示)

Figure 1　An example of knowledge graph (ovals and rectangles represent entities/classes and literals, respectively)

文将知识图谱内某个特定实体出现的所有三元组构成的集合称为该实体的实体描述. 例如, 图 1 呈现的一个简单的知识图谱例子, 实体 `Tim Berners-Lee` 的实体描述包含 10 个三元组, 分别表示了其与 `Weaving the Web` 和 `London` 等实体之间的关系, 以及一些类型和字面量信息.

知识图谱的一类应用是将实体相关的知识提供给用户阅读. 在大规模的知识图谱中, 描述一个实体的三元组可达上百条, 例如 DBpedia 中的实体 `Tim Berners-Lee`[1)]. 然而, 一些应用所提供的展示空间有限, 例如 Google Search 检索结果页面的实体卡片; 而将未经筛选的大量数据直接提供给用户, 也会增加用户获取关键信息的难度, 为用户阅读带来负担. 因此, 通常将实体描述的一个经过筛选的子集作为摘要提供给用户. 在开放域搜索和浏览应用中, 人工制定摘要的方式代价高耗时长且缺乏通用性. 为解决这些问题, 设计实体摘要的自动生成方法成为研究领域关注的问题.

实体摘要的目的是自动生成一个内容简短且包含关键信息的摘要以代替完整的实体描述. 在避免用户信息过载的同时, 能够基本满足用户的信息需求. 实体摘要可用于辅助用户完成各种任务, 一些现有工作研究面向特定任务的实体摘要方法, 例如, 信息检索中基于查询的实体摘要 [1~3], 面向实体链接任务的基于文档内容的实体摘要 [4], 以及用于共指消解任务的实体摘要 [5] 等. 与这些工作不同, 本文关注通用场景下的实体摘要问题, 即根据实体和知识图谱自身的特性, 自动生成可广泛应用于多种领域和应用的实体摘要.

目前已有一些研究工作 [6~10] 提出了通用场景下的实体摘要方法. 本文提出一种新的实体摘要方法, 将实体摘要问题建模为二次背包问题, 考虑了三元组内容在图结构中的重要性, 并从逻辑冗余、数值冗余和文本冗余 3 个方面量化三元组之间的冗余性. 此外, 考虑到知识图谱内容和面向应用的多样, 实体描述中可能存在较为面向技术而不易于用户阅读的内容 (例如 DBpedia 中属性 `oclc` 相关的三元组, 或 LinkedMDB 中的 `filmid`), 应当在摘要生成过程中予以过滤. 为此, 本文提出基于 N 元语法的熟知度指标, 利用文本语料度量三元组的可读性, 在实体摘要的生成过程中对低可读性的三元组予以惩罚. 综合考虑三元组的重要性, 可读性和摘要的冗余性, 本文将实体摘要问题建模为二次背包问题进行求解. 本文在两个公开数据集 Entity Summarization BenchMark (ESBM) 和 FACES Evaluation Dataset (FED) 上进行了实验, 对来自不同知识图谱的实体生成摘要, 并与 6 种现有实体摘要方法进行对比, 分析结果表明本文综合 3 种技术要素构建的实体摘要方法在各实验中均达到了当前最佳水平.

本文剩余部分将组织如下: 第 2 节先给出知识图谱和实体摘要问题的相关概念. 第 3 节介绍本文提出的基于组合优化模型的实体摘要方法 ESSTER. 第 4 节描述实验配置, 分析实验结果. 第 5 节讨论相关工作. 最后, 第 6 节总结全文并展望未来工作.

1) http://dbpedia.org/resource/Tim_Berners-Lee.

2 问题描述

知识图谱采用 RDF 数据模型来描述现实世界中的知识. RDF 用国际化资源标识符 (internationalized resource identifier, IRI) 标记现实中的实体, 并将实体相关的事实表示为 ⟨ 主语, 谓词, 宾语 ⟩ 的三元组形式. RDF 数据可表示为带标记的有向图, 三元组的主语和宾语作为图上节点的标记, 而谓词作为边的标记. 对于一个知识图谱 G, 设 $\mathbb{E}$ 为所有实体构成的集合, $\mathbb{C}$ 为实体类型集, $\mathbb{L}$ 为字面量集, $\mathbb{P}$ 为属性集. 则知识图谱中所有三元组的集合 T 可表示为 $T \subseteq \mathbb{E} \times \mathbb{P} \times (\mathbb{E} \cup \mathbb{C} \cup \mathbb{L})$. 其中, 属性集 $\mathbb{P}$ 又可划分为互不相交的 3 类: $\mathbb{R}$ 表示实体与实体间的关系 (如图 1 中属性 `author`), `TYPE` 用于关联实体与实体所属的类型 (如图 1 中属性 `type`), $\mathbb{A}$ 用于关联实体与字面量 (如图 1 中属性 `name` 和 `num of pages`). 因此三元组集合 T 又可表示为

$$T \subseteq (\mathbb{E} \times \mathbb{R} \times \mathbb{E}) \cup (\mathbb{E} \times \{\texttt{TYPE}\} \times \mathbb{C}) \cup (\mathbb{E} \times \mathbb{A} \times \mathbb{L}).$$

将一个三元组 $t \in T$ 的主语, 谓词, 宾语, 分别标记为 $\mathrm{subj}(t)$, $\mathrm{pred}(t)$, $\mathrm{obj}(t)$. 实体可出现在主语或宾语的位置. 对于一个待摘要的目标实体 $e \in \mathbb{E}$, 将与其直接关联的三元组所构成的集合称为该实体的实体描述 $\mathrm{desc}(e)$, 即

$$\mathrm{desc}(e) = \{t \in T : \mathrm{subj}(t) = e \text{ or } \mathrm{obj}(t) = e\}. \tag{1}$$

例如, 图 1 中实体 `Tim Berners-Lee` 的实体描述包含 10 个三元组. 从目标实体的角度, 实体描述中的一个三元组 $t \in \mathrm{desc}(e)$ 表示了目标实体 e 的属性 $\mathrm{prop}(t)$ 以及该属性的取值 $\mathrm{val}(t)$, 称为 "属性 – 值对", 即 $\langle \mathrm{prop}(t), \mathrm{val}(t) \rangle$:

$$\mathrm{prop}(t) = \begin{cases} \mathrm{pred}(t), & \text{if } \mathrm{subj}(t) = e, \\ \mathrm{pred}(t)^-, & \text{if } \mathrm{obj}(t) = e, \end{cases} \quad \mathrm{val}(t) = \begin{cases} \mathrm{obj}(t), & \text{if } \mathrm{subj}(t) = e, \\ \mathrm{subj}(t), & \text{if } \mathrm{obj}(t) = e, \end{cases} \tag{2}$$

其中 $\mathrm{pred}(t)^-$ 表示谓词 $\mathrm{pred}(t)$ 的反向含义, 即三元组 $\langle e_1, p, e_2 \rangle$ 与 $\langle e_2, p^-, e_1 \rangle$ 表达相同的事实. 因此, 三元组 $t \in \mathrm{desc}(e)$ 可统一改写为 $\langle e, \mathrm{prop}(t), \mathrm{value}(t) \rangle$. 由于目标实体相同, 实体描述内三元组之间的比较可以视为相应的属性 – 值对之间的比较.

实体摘要问题, 即从目标实体的实体描述中选出一个满足容量限制的子集 S 作为摘要, 以尽可能满足用户的信息需求. 在以往工作 [6~9, 11] 中, 容量限制被设定为摘要 S 中所含三元组的个数, 记为 k. 假定对摘要 S 的质量评价为 $\mathrm{score}(S)$, 则实体摘要问题可表示为

$$\text{find} \quad \operatorname*{arg\,max}_{S \subseteq \mathrm{desc}(e)} \mathrm{score}(S) \quad \text{s.t. } |S| \leqslant k. \tag{3}$$

实体摘要问题的关键是定义 score 函数. 本文认为一个高质量的摘要应具备 3 种性质: 高重要性、高可读性、低冗余性.

3 ESSTER 方法

本节详细介绍本文提出的实体摘要方法 ESSTER. 先阐述实体摘要任务中对高重要性、高可读性和低冗余性这 3 种性质的要求与本文的解决思路, 然后分别对这 3 种性质提出相应的量化方式, 最后利用组合优化模型综合这三者建模并求解.

3.1 概述

实体摘要方法 ESSTER 的设计初衷为生成具备高重要性、高可读性和低冗余性的摘要. 本小节对这 3 种性质及其要求分别阐述如下.

高重要性. 即选入摘要的各三元组自身包含重要的信息. 属性 – 值对在知识图谱中的自信息常作为三元组所含信息的一种度量, 认为一个属性 – 值对出现在越少的实体描述中, 则它一旦出现所提供的信息就越多, 越有助于区分目标实体与其他实体. 然而, 不同知识图谱中适合作为摘要的三元组表现的特征不同, 三元组的重要性度量应该适应不同知识图谱的差异. 为此, 本文提出对属性和取值分别考虑, 基于其在全局和局部图结构中的特征度量对不同现象的偏好, 并通过组合这些不同偏好以适应不同知识图谱在属性、取值分布上的差异, 提高方法的通用性.

高可读性. 实体摘要的主要应用之一是提供给终端用户阅读, 辅助用户理解目标实体. 生僻和过于技术性的内容将增加用户的阅读难度, 降低实体摘要的有用性. 而知识图谱在构造时为了适应不同应用的需求, 会包含各类数据, 包括易于程序处理却不适合人类阅读的编码、代号等数据. 高质量的摘要应该尽可能提供用户易于理解的内容, 因此, 实体摘要方法应该具备过滤低可读性三元组的能力. 现有实体摘要工作尚未关注到这一问题. 本文提出借助文本语料度量三元组的可读性, 并将其作为对三元组质量的赋权, 用于影响对三元组的筛选过程.

低冗余性. 即避免携带重复内容的三元组被同时选入摘要, 以使生成的摘要在容量限制下能包含尽可能多样的内容. 降低冗余的关键在于如何判断三元组之间存在的重复关系, 以及设置对冗余的不同容忍程度. 简单的基于规则的强约束可能导致最终生成的摘要规模低于容量限制. 本文结合逻辑推理, 数值比较和文本相似度对两个三元组之间的冗余程度打分, 将该打分作为摘要质量评价的惩罚项来降低冗余, 以灵活适应对不同组合条件下的三元组选择.

3.2 结构重要性

在现有实体摘要方法中, 基于三元组在知识图谱结构中的统计特性度量三元组的重要性已有多种实现, 其中基于三元组自信息的方式最为常见 [6, 8∼10], 偏好于选择属性 – 值对在全局图结构中较为少见的三元组, 或者结合取值的流行度 [8, 9, 11], 偏好于选择取值较为常见的三元组. 而这些单一的偏好对特性不同的数据集的适应能力有限. 本文提出的方法同时考虑三元组的属性和取值在全局和局部图结构中的特征, 对不同特征加以区分, 并组合对这些特征的不同偏好以适应不同知识图谱的特性.

首先, 属性在整个知识图谱结构中的流行度反映了其通用性, 体现了该知识图谱中普遍强调的重要含义, 例如 DBpedia 中的属性 `label`, `type`, LinkedMDB 中的 `genre`, `language`, 应该考虑选入摘要. 对于具备这样属性的三元组, 若其取值也极为常见, 则难以将当前实体与知识图谱中其他实体进行区分, 例如 $\langle$`type`, `Thing`$\rangle$. 而相对低流行度的取值则能够体现当前实体的独特性, 例如 $\langle$`type`, `Scientist`$\rangle$. 令 $\mathrm{ppop}_{\mathrm{global}}(t) \in [0,1]$ 为三元组 t 的属性在知识图谱中的全局流行度, $\mathrm{vpop}(t) \in [0,1]$ 为其取值的流行度, 则可将这类度量表示为

$$\mathrm{ch}(t) = \mathrm{ppop}_{\mathrm{global}}(t) \cdot (1 - \mathrm{vpop}(t)). \tag{4}$$

此外, 多值属性的存在可能导致对三元组的排序被带有单一属性的三元组主导 (例如 DBpedia 中的 `type`, `subject`, `influenced` 等属性会在一个实体描述中出现几十次). 从提高实体摘要内容多样性的角度考虑, 本文根据属性在局部图结构中的流行度区分多值属性, 并给予一定惩罚. 在此基础上, 为了避免选入过于偏技术应用的三元组 (如图 1 中属性 `PND`, `isbn`), 对取值流行度高的取值予以奖励. 令

$\mathrm{ppop}_{\mathrm{local}}(t) \in [0,1]$ 为三元组 t 的属性在实体描述所对应局部图结构中的流行度, 则可表示为

$$\mathrm{div}(t) = (1 - \mathrm{ppop}_{\mathrm{local}}(t)) \cdot \mathrm{vpop}(t). \tag{5}$$

引入参数 $\alpha \in [0,1]$ 来调节不同数据集对这两种现象的不同偏好, 则一个三元组的重要性打分为

$$W_{\mathrm{struct}}(t) = \alpha \cdot ch(t) + (1-\alpha) \cdot \mathrm{div}(t). \tag{6}$$

在实现中, 对 $\mathrm{ppop}_{\mathrm{global}}(t), \mathrm{ppop}_{\mathrm{local}}(t), \mathrm{vpop}(t)$ 的估计来源于属性和值在知识图谱图结构中的出现频度, 即三元组 t 的属性 $\mathrm{prop}(t)$ 在知识图谱中出现在不同实体描述中的次数 $\mathrm{pfreq}_{\mathrm{global}}(t)$, 属性 $\mathrm{prop}(t)$ 在目标实体的实体描述 $\mathrm{desc}(e)$ 中出现的次数 $\mathrm{pfreq}_{\mathrm{local}}(t)$, 以及取值 $\mathrm{val}(t)$ 在知识图谱所有三元组中出现的次数 $\mathrm{vfreq}(t)$. 将这些频度值归一化到 $[0,1]$ 区间, 即可获得相应的流行度值, 即

$$\begin{aligned}
&\mathrm{pfreq}_{\mathrm{global}}(t) = |\{e' \in \mathbb{E} : \exists t' \in \mathrm{desc}(e'), \mathrm{prop}(t') = \mathrm{prop}(t)\}|, \quad \mathrm{ppop}_{\mathrm{global}}(t) = \frac{\log(\mathrm{pfreq}_{\mathrm{global}}(t)+1)}{\log(\|\mathbb{E}\|+1)},\\
&\mathrm{pfreq}_{\mathrm{local}}(t) = |\{t' \in \mathrm{desc}(e) : \mathrm{prop}(t') = \mathrm{prop}(t)\}|, \qquad \mathrm{ppop}_{\mathrm{local}}(t) = \frac{\log(\mathrm{pfreq}_{\mathrm{local}}(t)+1)}{\log(\|\mathrm{desc}(e)\|+1)},\\
&\mathrm{vfreq}(t) = |\{t' \in T : \mathrm{val}(t') = \mathrm{val}(t)\}|, \qquad \mathrm{vpop}(t) = \frac{\log(\mathrm{vfreq}(t)+1)}{\log(\|T\|+1)}.
\end{aligned} \tag{7}$$

值得讨论的是, 重要性在不同领域、不同任务下有不同标准. 在一些特定场景中, 重要性的度量需要结合与上下文的相关度等因素. 而本文所关注的通用场景下的实体摘要, 并不假定存在这类上下文信息, 因此更关注三元组在知识图谱结构中的表现. 基于知识图谱结构的重要性度量可采用多种不同的计算方式, 本文考虑的结构重要性用参数化的方式综合了三元组在知识图谱全局和局部图结构中的表现, 从而能表达重要性的复杂内容, 并适应不同知识图谱的结构特点.

3.3 文本可读性

图 1 提供了实体 `Tim Berners-Lee` 的属性 `PND` 的信息. 该属性名为德语 Personennamendatei 的缩写, 其取值表示由德国国家图书馆发布的用于唯一标识每个人的编号. 然而, 除非借助外部资源, 普通用户难以将 `PND` 编号与特定的人物特性联系起来, 甚至不了解 `PND` 属性本身的含义. 相比之下, `author` 和 `birth place` 等日常属性则更易理解. 这说明对三元组的理解需要借助不同程度的背景知识, 某些属性可能对特定用户有用 (例如图书管理员可能对 `PND` 非常了解), 却不易于非专家型一般用户的理解. 因此, 在通用场景中为普通用户生成摘要, 需要考虑三元组的可读性.

为了比较不同三元组在用户阅读体验上的上述差异, 本文考虑对可读性进行量化, 基于属性 $\mathrm{prop}(t)$ 的文本来计算三元组的可读性 $Q(t)$. 正如上述例子所表现的, 理解不同属性需要不同程度的知识储备. 为此, 假设在用户日常阅读环境中越常见的文本越为用户所熟知, 因而越容易被理解. 本文以开放域文本语料模拟用户日常阅读环境. 设语料中包含 B 个文档 (如网页, 书籍), 用 $b(t)$ 表示属性 $\mathrm{prop}(t)$ 的文本在该语料中出现的文档数, $m(t)$ 表示这 $b(t)$ 个文档中用户曾观察到的文档数量. 则三元组的可读性 $Q(t)$ 是关于 $m(t)$ 的函数, 定义为

$$Q(t) = \mathrm{familiarity}(m(t)), \tag{8}$$

其中 familiarity 为关于 $m(t)$ 的非递减函数, 取值在 $[0,1]$ 区间. 用次线性函数定义 familiarity 为

$$\mathrm{familiarity}(m(t)) = \frac{\log(m(t)+1)}{\log(B+1)}. \tag{9}$$

在实际实现中, 通常难以获取每个用户的 $m(t)$ 值, 因此改为计算 familiarity($m(t)$) 对所有 $m(t)$ 可能值的期望值. 即, 令 M 为该语料中用户已读的文档数量, 将式 (8) 改写为

$$Q(t)=\sum_{m=0}^{\min(b(t),M)}\frac{\binom{b(t)}{m}\cdot\binom{B-b(t)}{M-m}}{\binom{B}{M}}\cdot\text{familiarity}(m). \tag{10}$$

为简化计算, 将 M 定为常量, 可作为参数进行调节或根据先验知识直接设定. 由此, 则上述公式可在常数时间内计算. 特别地, 某些属性的文本可能从未在语料中出现过, 从而导致 $b(t)=0$, 进而 $Q(t)=0$. 为避免出现零概率的情况, 采取平滑策略, 在语料中加入伪文档并假定该文档包含所有属性的文本. 即式 (10) 中的 B 和 $b(t)$ 值均增 1, 以确保 $Q(t)>0$.

本文在实现中采用 Google Books Ngram[2)]作为语料. 该语料提供书籍中出现文本的统计量而非原始文档, 例如各 n-gram 在语料中出现的文档数量, 其中 n-gram 的长度 n 在 $[1,5]$ 范围内. 这就导致对于具有长文本的属性, 若其文本长度超过 n 的最大值 5, 则无法准确获知其 $b(t)$ 值. 此外, 随着文本长度的增大, 其对应的 n-gram 数据越稀疏. 因此采用其上界来近似 $b(t)$ 的值, 令 n_T 为一指定长度, 用属性文本的 n_T-gram 来计算该上界. 具体而言, 若一个属性文本包含 w 个词 ($w>n_T$), 则含有 $(w-n_T+1)$ 个 n_T-gram. 对其中的每个 n_T-gram, 从语料提供的统计数据中获取其出现的文档数量, 并取各值的最小值作为对 $b(f)$ 的近似, 该最小值即为 $b(f)$ 真实值的上界. 本文实验中取 $n_T=2$.

在实际使用中, $Q(t)$ 可作为三元组打分的辅助权重. 为避免 $Q(t)$ 取值分布的过度倾斜带来的过度惩罚, 采用对数函数对其值进行调整, 即, 最终文本可读性权重为

$$W_{\text{text}}(t)=\log(Q(t)+1). \tag{11}$$

需要指出的是, 虽然高可读性与高重要性都存在对频度的偏好, 但两者的要求并不矛盾. 首先, 两者所考虑的频度, 是对三元组的不同信息在两种不同环境下做出的统计. 重要性考虑的是三元组在知识图谱中的结构特点, 关注于三元组各元在知识图谱中的使用频度所体现的独特性、通用性; 而可读性则基于文本语料中统计的频度, 体现三元组文本在用户的语言环境中被熟知的程度. 此外, 重要性并不单纯偏好低频取值, 而是考虑了属性和取值在全局和局部图谱中的综合表现. 对高可读性的考虑是对高重要性的一种补充和平衡, 在重要性相同的三元组之间, 更优先选择在语料中较为常见的文本作为摘要内容.

3.4 冗余度

在图 1 中, 实体 `Tim Berners-Lee` 的 `name`, `surname` 和 `given name` 属性对应的三元组的内容存在大量重复. 相比于等量的不存在重复的三元组 (例如将 `surname` 和 `given name` 的三元组替换为 `birth place`, `author` 的三元组), 这样具备重复信息的三元组集所能提供的有用信息更少. `Tim Berners-Lee` 的 `type` 属性对应的两个三元组分别描述了该实体属于 `Scientist` 类型和 `Person` 类型, 而 `Scientist` 与 `Person` 之间存在上下位关系, 后者对应三元组的内容可由前者推理得到, 因为一个实体属于 `Scientist` 则必然也属于 `Person`, 因此也可认为两者内容存在重复. 上述现象表明, 三元组之间存在重复内容, 带来信息冗余. 在有限的容量下, 为包含多样的内容, 存在冗余的三元组不希望被同时选中. 若仅考虑对单个三元组的打分, 这些内容相似的三元组的得分也可能较为相近, 从而很可能被连续选入摘要中. 为了避免存在冗余的三元组被同时选入摘要, 需要对三元组间的冗余度进行量化, 以在摘要生成过程中根据已选情况动态调整对不同三元组的偏好.

2) https://books.google.com/ngrams.

本文将实体摘要的冗余度转化为对摘要中所含三元组两两之间冗余度的计算. 对于任意两个三元组, 综合考虑其逻辑冗余、数值冗余、文本冗余 3 个方面计算其冗余程度. 计算两三元组间冗余度 $\mathrm{ovlp}(t_i, t_j)$ 的完整过程见算法 1.

算法 1 Redundancy

Input: $t_i, t_j \in \mathrm{desc}(e)$;
Output: $\mathrm{ovlp}(t_i, t_j)$;

1: **if** $\mathrm{prop}(t_i) = $ `rdf:type` and $\mathrm{prop}(t_j) = $ `rdf:type` and (`subClassOf`$(\mathrm{val}(t_i), \mathrm{val}(t_j))$ or `subClassOf`$(\mathrm{val}(t_j), \mathrm{val}(t_i))$) **then**
2: $\quad \mathrm{ovlp}(t_i, t_j) \Leftarrow 1$;
3: **else if** $\mathrm{val}(t_i) = \mathrm{val}(t_j)$ and (`subPropertyOf`$(\mathrm{prop}(t_i), \mathrm{prop}(t_j))$ or `subPropertyOf`$(\mathrm{prop}(t_j), \mathrm{prop}(t_i))$) **then**
4: $\quad \mathrm{ovlp}(t_i, t_j) \Leftarrow 1$;
5: **else**
6: $\quad \mathrm{sim}_p(t_i, t_j) = $ `ISub`$(\mathrm{prop}(t_i), \mathrm{prop}(t_j))$;
7: $\quad$ **if** `isNumber`$(\mathrm{val}(t_i))$ and `isNumber`$(\mathrm{val}(t_j))$ **then**
8: $\quad\quad$ **if** $\mathrm{val}(t_i) = \mathrm{val}(t_j)$ **then**
9: $\quad\quad\quad \mathrm{sim}_v(t_i, t_j) \Leftarrow 1$;
10: $\quad\quad$ **else if** $\mathrm{val}(t_i) \cdot \mathrm{val}(t_j) \leqslant 0$ **then**
11: $\quad\quad\quad \mathrm{sim}_v(t_i, t_j) \Leftarrow -1$;
12: $\quad\quad$ **else**
13: $\quad\quad\quad \mathrm{sim}_v(t_i, t_j) \Leftarrow \frac{\min\{\|\mathrm{val}(t_i)\|, \|\mathrm{val}(t_j)\|\}}{\max\{\|\mathrm{val}(t_i)\|, \|\mathrm{val}(t_j)\|\}}$;
14: $\quad\quad$ **end if**
15: $\quad$ **else**
16: $\quad\quad \mathrm{sim}_v(t_i, t_j) \Leftarrow$ `ISub`$(\mathrm{val}(t_i), \mathrm{val}(t_j))$;
17: $\quad$ **end if**
18: $\quad \mathrm{ovlp}(t_i, t_j) \Leftarrow \max\{\mathrm{sim}_p(t_i, t_j), \mathrm{sim}_v(t_i, t_j), 0\}$;
19: **end if**

首先, 利用本体知识来判断逻辑冗余. 本体知识包含属性和类型之间的上下位关系 (`subClassOf`, `subPropertyOf`). 实体描述中, 属性为 `rdf:type` 的三元组描述了实体的类型归属情况, 若两个三元组所描述的实体类型存在上下位关系, 则其一可由另一个推理得出, 因此可判定这两个三元组间存在逻辑冗余. 类似的, 当两个三元组取值相同, 而属性之间存在上下位关系, 同样认为两者之间存在逻辑冗余. 即算法 1 中第 1∼4 行.

对于其他情况, 分别计算三元组属性之间、取值之间的相似度来度量其冗余程度. 首先, 计算两个三元组的属性之间的相似度 $\mathrm{sim}_p(t_i, t_j) \in [-1, 1]$. 本文通过属性的 `rdfs:label` 信息获取属性的文本形式, 再采用字符串相似度指标 ISub [12] 计算两属性文本间的相似度. 即算法 1 中第 6 行. 然后, 计算两个三元组的取值之间的相似度 $\mathrm{sim}_v(t_i, t_j) \in [-1, 1]$. 对于两三元组取值同为数值的情况, 根据数值的大小分 3 种情况判断两者的相似性, 即算法 1 中第 7∼14 行. 在取值并非同为数值的情况下, 两个三元组取值之间的相似度为其文本字符串间的相似度. 若取值为实体或类型, 通过该取值的 `rdfs:label` 信息获得取值的文本形式; 若取值为字面量, 则其文本形式即其自身内容. 将两个取值 $\mathrm{val}(t_i)$ 和 $\mathrm{val}(t_j)$ 的文本形式视为字符串, 采用字符串相似度指标 ISub [12] 计算两者的相似度, 即算法 1 中第 16 行. 综合属性与取值的相似度, 取其最大值以使冗余度对属性 – 值对的任一部分存在相似敏感. 如算法 1 中第 18 行, 即求得基于文本和数值计算的冗余度.

最终, 可得两个三元组之间的冗余度 $\mathrm{ovlp}(t_i, t_j)$, 其取值在 $[0, 1]$ 区间, 值越大则表明两者存在越大程度的冗余. 实体摘要的冗余度为其所含三元组之间成对冗余度之和.

3.5 组合优化模型

通过上述讨论, 对结构重要性、文本可读性和冗余性都提出了量化方式. 为了生成具备高重要性、高可读性和低冗余性的实体摘要, 可将实体摘要 S 的质量打分定义为

$$\mathrm{score}(S)=(1-\delta)\cdot\sum_{t\in S}[W_{\mathrm{struct}}(t)+W_{\mathrm{text}}(t)]+\delta\cdot\sum_{t_i,t_j\in S}-\mathrm{ovlp}(t_i,t_j),\tag{12}$$

其中 $\delta\in[0,1]$ 为待调参数, 用于调节实体摘要对冗余的容忍程度. 在给定容量限制 k 下, 以最大化上述质量打分为优化目标的实体摘要问题可建模为二次背包问题 (quadratic knapsack problem, QKP) [13], 实体描述 $\mathrm{desc}(e)$ 中的每个三元组可视为背包问题中的一个物品. 将 $\mathrm{desc}(e)$ 内各三元组依次编号为 t_1, t_2 至 $t_{\|\mathrm{desc}(e)\|}$, 引入指示变量 x_i $(i=1,\ldots,\|\mathrm{desc}(e)\|)$, 以指示三元组 t_i 是否被选入摘要, 则模型可形式化为

$$\begin{aligned}&\text{maximize}\quad\sum_{i=1}^{\|\mathrm{desc}(e)\|}\sum_{j=i}^{\|\mathrm{desc}(e)\|}\mathrm{profit}_{i,j}\cdot x_i\cdot x_j,\\&\text{s.t.}\quad\sum_{i=1}^{\|\mathrm{desc}(e)\|}x_i\leqslant k,\quad x_i\in\{0,1\},\ \text{for } i=1,\ldots,\|\mathrm{desc}(e)\|,\end{aligned}\tag{13}$$

其中 $\mathrm{profit}_{i,j}$ 为背包中同时选入第 i 和 j 个物品带来的收益. 参照式 (12) 对摘要质量的打分, 则 $\mathrm{profit}_{i,j}$ 可定义为

$$\mathrm{profit}_{i,j}=\begin{cases}(1-\delta)\cdot[W_{\mathrm{struct}}(t_i)+W_{\mathrm{text}}(t_i)], & \text{if } i=j,\\ \delta\cdot(-\mathrm{ovlp}(t_i,t_j)), & \text{otherwise.}\end{cases}\tag{14}$$

二次背包问题对应的决策问题属于 NP- 完全问题, 目前仅存在基于动态规划的伪多项式时间算法及一些启发式算法. 本文的实现中采用了 Yang 等 [14] 提出的启发式算法.

4 实验与分析

本节实验评价本文所提出的实体摘要方法. 该实验在两个公开数据集上将本文方法与 6 种现有实体摘要方法进行比较, 并分别分析了方法中的 3 种度量对实体摘要任务的影响.

4.1 实验配置

4.1.1 实验数据

本实验采用实体摘要领域两个常用的公开数据集 ESBM[3)]和 FED[4)], 并设定摘要容量限制 $k=5$.

ESBM v1.1 数据集. 为目前实体摘要领域可公开获取的规模最大的数据集, 其包含的实体来自两个不同的知识图谱: DBpedia[5)]和 LinkedMDB[6)]. DBpedia 为百科型知识图谱, 其实体描述内容来自于大众编辑的在线百科 Wikipedia[7)], 包含的实体类型多样. LinkedMDB 则为特定领域的知识图谱, 包含影视领域的人物、作品等类型实体. 两个数据集的实体类型、属性分布均有较大差异, 给实体摘要方

3) https://w3id.org/esbm.
4) http://wiki.knoesis.org/index.php/FACES.
5) http://dbpedia.org.
6) http://linkedmdb.org.
7) https://www.wikipedia.org.

法带来不同的挑战. 具备良好通用性的实体摘要方法, 应该对这两个知识图谱的实体都能生成高质量的实体摘要. ESBM v1.1 为 175 个实体人工标注了 1050 个以 $k=5$ 为容量限制的实体摘要, 其中 125 个实体来自 DBpedia 2015-10 (记为 ESBM-D), 50 个来自 LinkedMDB (记为 ESBM-L). 每个实体对应由 6 个不同用户提供的理想实体摘要.

FED 数据集. 由 Gunaratna 等 [8] 构造, 该数据集包含 50 个来自 DBpedia 3.9 的实体, 其实体描述所考虑的属性集与 ESBM-D 有所不同, 且仅包含描述两实体间关系的三元组. 该数据集的每个实体由 6 至 8 个用户标注理想实体摘要, 最终可形成容量限制 $k=5$ 下的 373 个实体摘要.

为避免实验过程中的偶然性, 本文采用 5 折交叉验证的方式对模型调参并统计结果. 即各数据集按实体划分为互不相交的 5 等份, 组织成 5 折的训练和测试集, 每折训练和测试集合包含的实体数量比例为 4:1. 在训练集上对方法调参, 将训练集上的最佳参数用在测试集上, 并记录测试集上的评估结果, 5 折测试集上的平均值作为方法的最终得分.

4.1.2 评价指标

给定一个实体摘要, 其质量评价来自于与相应目标实体的各理想摘要的比较. 本文依照 ESBM v1.1 的做法, 采用 F-measure 来度量摘要质量, 即假定 S 为待评价摘要, S^* 为一个理想摘要, 则

$$\text{Precision}=\frac{\|S\cap S^*\|}{\|S\|},\quad \text{Recall}=\frac{\|S\cap S^*\|}{\|S^*\|},\quad \text{F-measure}=\frac{2\cdot\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}}. \tag{15}$$

摘要 S 的 F-measure 为其与目标实体各理想摘要分别计算 F-measure 的平均值. 一个实体摘要系统的质量则是其为测试集各实体生成的摘要 F-measure 的平均值.

4.1.3 对比方法

本实验将 ESSTER 与 6 种现有实体摘要方法进行对比, 其实现与配置方式介绍如下.

RELIN [6] 是最早提出的实体摘要方法之一, 采用带权 PageRank 模型, 综合考虑了三元组的自信息和三元组之间的相关度. 由于其原本用到的 Google search API 不再免费提供支持, 本文在实现中将其原本基于 Google 检索的相关度度量替换为字符串相似度指标 ISub [12].

DIVERSUM [7] 提出基于对属性的强约束构造具备多样性的实体摘要. 其 witness count 指标在本文所使用的数据集上无对应的概念, 因此在本实现中并未考虑.

FACES [8] 及其扩展方法 FACES-E [9] 基于对三元组的聚类生成具备多样性的摘要. 其中 FACES-E 是 ESBM v1.1 公布的评测结果中 ESBM-L 上表现最佳的方法. 本实验采用从其作者处获得的源码和配置. 由于 FACES-E 所采用的 UMBC SimService 不再提供服务, 改用 ISub [12] 替代.

CD [10] 也是 ESBM v1.1 上效果最佳的方法之一, 综合考虑三元组的独特性与三元组之间的冗余度. 本实验中 CD 的实现采用了从原作者处获得的代码.

LinkSUM [11] 关注于处理取值为实体的三元组. 其先根据三元组取值的 PageRank 和 backlink 指标对取值做打分排序, 再筛选取值相同的三元组的属性. 对来自 LinkedMDB 的实体, 本文通过其与 DBpedia 实体间的对应关系获取 backlink 数据.

ESSTER 即本文提出的方法, 本文实验中设定式 (10) 中的 M 值为 40, 而式 (6) 中的参数 α 和式 (14) 中的参数 δ 则在 5- 折划分各折的训练集上在 $[0,1]$ 区间内以 0.01 为步长调参. 在消融实验中, 分别将 3 个部分的度量值置零以观察其对摘要结果的影响, 删除对应部分后的方法分别记为 ESSTER-S, ESSTER-T 和 ESSTER-R.

表 1　各实体摘要方法的 F-measure 结果

Table 1　F-measure of entity summarizers[a)]

	ESBM-D	ESBM-L	FED
RELIN	0.242 (0.120)	0.203 (0.125)	0.127 (0.085)
DIVERSUM	0.249 (0.136)	0.207 (0.127)	0.112 (0.078)
FACES	0.270 (0.144)	0.169 (0.085)	0.145 (0.089)
FACES-E	0.280 (0.142)	0.313 (0.116)	0.145 (0.089)
CD	0.283 (0.134)	0.217 (0.101)	0.136 (0.076)
LinkSUM	0.287 (0.132)	0.140 (0.101)	0.239 (0.121)
ESSTER	0.305 (0.132) ▲▲▲△○○	0.347 (0.077) ▲▲▲△▲▲	0.229 (0.118) ▲▲▲▲▲○

a) Significant improvements of ESSTER over each baseline are indicated by ▲ ($p < 0.01$) and △ ($p < 0.05$). Insignificant differences are indicated by ○.

表 2　消融实验的 F-measure 结果

Table 2　F-measure of ablation study

	ESBM-D			ESBM-L			FED		
	Mean	diff	p	Mean	diff	p	Mean	diff	p
ESSTER	0.305	–	–	0.347	–	–	0.229	–	–
ESSTER-S	0.264	−0.041	0.000	0.247	−0.101	0.000	0.140	−0.089	0.000
ESSTER-T	0.298	−0.007	0.489	0.305	−0.042	0.001	0.218	−0.011	0.167
ESSTER-R	0.222	−0.083	0.000	0.325	−0.022	0.025	0.211	−0.019	0.042

特别地, RELIN, CD 和 LinkSUM 均需要对两部分打分的权重参数做调节, 而调参结果会受训练 – 测试集划分的影响, 用 ESBM v1.1 网站上提供的结果与本文方法直接比较显得不公平. 因此, 本文采用与 ESSTER 同样的 5 折划分重新对这些方法进行调参和测试, 并将该结果与本文方法对比.

4.2　实验结果

表 1 给出了各方法在各数据集上结果的 F-measure 均值与标准差, 并采用双边配对样本 t 检验比较 ESSTER 与各对比方法差异的显著性. 与以往实体摘要方法相比, ESSTER 在除 FED 外的所有配置下均高于现有工作的最佳效果. 在 ESBM-D 上, ESSTER 比对比方法的最佳效果提升 0.018; 而在 ESBM-L 上更是比最佳对比方法提升了 0.034, 显著高于所有对比方法 ($p > 0.05$). 在 FED 上, 大部分对比方法的效果均低于 0.15, 仅本文提出的 ESSTER 与 LinkSUM 的 F-measure 值超过 0.2, 且两者的差异并不显著. 而 LinkSUM 在该数据集上效果较好, 主要是由于 FED 数据集的实体描述仅包含用于描述实体与实体间关系的三元组, LinkSUM 的设计正是针对这类三元组, 而不考虑对取值为字面量的三元组的处理. 因此, 综合考虑方法对不同类型三元组的支持程度, 以及在各数据集上摘要的效果, ESSTER 都达到了最高水平.

为分析 ESSTER 中 3 种技术要素的有效性, 对 3 种要素进行消融实验, 在完整 ESSTER 方法的基础上分别去除 3 种要素, 即相应部分的打分置零. 表 2 列出了该实验的结果, 并给出消融版本与采用完整的 ESSTER 结果均值之间的差异 (diff) 和双边配对样本 t 检验的 p- 值结果. 可以看出, 删除各部分后实体摘要的 F-measure 均值均有所下降, 但在不同数据集上表现略有不同. 在 ESBM-D 和 FED 上, 删除结构重要性 (S) 和冗余度 (R) 对摘要效果的影响最为显著, 而在 ESBM-L 上, 3 个消融版本

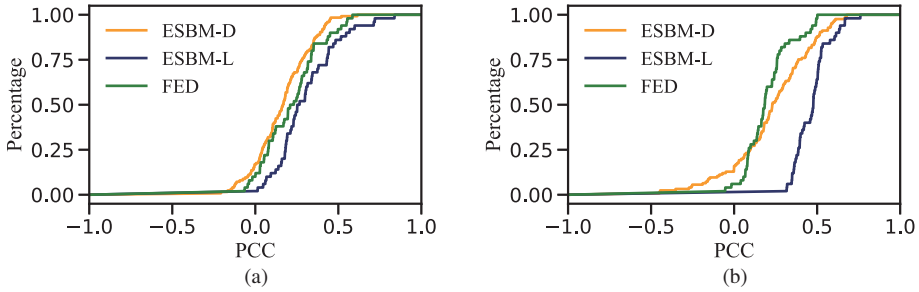

图 2 (网络版彩图) 打分方式与理想打分之间皮尔逊积矩相关系数的累计分布

Figure 2 (Color online) Cumulative distribution of Pearson correlation coefficient between weights and ideal importance scores. (a) W_{struct}; (b) W_{text}

均带来效果的显著下降 ($p < 0.05$). 本文接下来分别分析 3 种要素带来的影响.

4.2.1 重要性的影响

在实验数据中, 三元组被选入理想摘要的次数可视为综合多个用户意见对三元组的理想打分, 体现了用户对三元组的质量评价. 对每个实体, 计算实体描述内各三元组的结构重要性打分 W_{struct} 与用户理想打分之间的皮尔逊积矩相关系数 (Pearson correlation coefficient, PCC). 图 2(a) 呈现了各数据集上各实体 PCC 值的累计分布. PCC 的取值范围为 $[-1, 1]$, 用于度量两个随机变量间的线型相关性, PCC 值越接近 1 表明两个随机变量间正相关性越高, 越接近 -1 则表明存在负相关, 接近 0 值则表明两变量间几乎不存在线性相关性. 一个好的重要性打分应该与用户理想打分呈正相关性, 即在越多的实体上 PCC 结果越接近 1 越好. 图 2(a) 中的一个点 (x, y) 表示 PCC $\leqslant x$ 的实体数在该数据集中占比为 y. 在 ESBM-D, ESBM-L, FED 3 个数据集上, PCC 值为正的实体所占比例分别为 87%, 100% 和 96%, 且 ESBM-L 中 36% 的实体 PCC 值大于 0.5, 体现出较强的正相关性. 可以看出结构重要性打分 W_{struct} 在各数据集上都在大部分实体上表现出与用户理想打分的正相关性, 说明该打分方式与用户对摘要中三元组质量的评价较为一致.

4.2.2 可读性的影响

表 3 呈现了分别来自 ESBM-D, ESBM-L 和 FED 的属性中, 可读性得分最高和最低的前 10 个属性. 从中可以观察到, 可读性较高的属性通常由较为常见的单词构成, 且大多数仅包含 1 个单词. 低可读性属性中常见编号相关的属性 (如 `IMDB id`, `INSEE code`) 以及包含生僻单词的属性 (如 `państwo`), 可见采用可读性确实有助于过滤掉含这类属性的三元组.

分析每个实体摘要任务中, 各属性的熟知度与其所在三元组被选入理想摘要中的次数之间的 PCC, 该结果在各数据集上的累计分布如图 2(b) 所示. 在 ESBM-D 上, 82% 的实体上 PCC 值为正, FED 上 PCC 值为正的实体则达到 96%, 而 ESBM-L 则在所有实体上 PCC 值都为正, 且在 14% 的实体上呈现出高正相关性 (PCC > 0.5). 因此可以看出三元组的可读性对构造让用户满意的实体摘要可起到正面作用.

4.2.3 冗余性的影响

为分析实体摘要中冗余性的影响, 计算实体描述、理想摘要, 以及各实体摘要方法生成的摘要的总体冗余度, 即集合中所含三元组的成对冗余度之和. 各数据集上的结果见表 4.

表 3　各数据集上文本可读性最高和最低的前 10 个属性

Table 3　Top-10 properties with highest or lowest readability in each dataset

ESBM-D		ESBM-L		FED	
Highest	Lowest	Highest	Lowest	Highest	Lowest
time	draft year	made	link source	other	population blank1 title
long	debut team	subject	filmid	before	timezone DST
order	IMDB id	country	story contributor	after	computing media
number	type of tennis surface	date	film story contributor	years	demonym
course	siler medalist	language	music contributor	order	sovereignty type
name	UTC offset	type	director directorid	name	languages2 type
subject	NRHP reference number	page	director name	state	cctId
added	route type abbreviation	title	director nytimes id	parts	location country
result	serving railway line	writer	actor name	ground	państwo
position	bionomial authority	performance	actor Netflix id	country	flaglink

表 4　实体描述, 理想摘要及各摘要方法生成的摘要的冗余度

Table 4　Redundancy of entity description, ideal entity summary and summaries generated by entity summarizers

	ESBM-D	ESBM-L	FED
Desc	203.69	431.99	140.78
Ideal	1.04	1.84	0.89
RELIN	3.04	3.45	2.22
DIVERSUM	0.39	1.29	1.64
FACES	0.75	0.30	1.05
FACES-E	1.29	0.76	1.05
CD	0.02	0.00	0.00
LinkSUM	2.45	4.72	1.47
ESSTER	0.02	1.17	1.69

从表 4 中可看出实体描述内存在大量冗余现象, 其冗余度远高于各摘要. FED 数据集内实体摘要的冗余度最小, 而 ESBM-L 实体的平均冗余度达到 ESBM-D 实体的两倍. 实体摘要的冗余度大幅小于实体描述的冗余度, 也体现了实体摘要比实体描述更为精简的特点. 各数据集理想摘要的冗余度则较为接近, 且值都在 2 以内.

各实体摘要方法中, RELIN 未考虑去冗余, LinkSUM 对冗余的要求也较为简单, 两者生成的实体摘要的冗余度也都明显高于其他方法. 其他几个考虑了冗余的方法, 生成的摘要的冗余度都较为接近. 在 ESBM-D 实体上, CD, ESSTER 和 DIVERSUM 的冗余度最低, 而在 ESBM-L 和 FED 上则是 CD, FACES 和 FACES-E 获得最低冗余度. 说明这些摘要方法对冗余度的控制均得到了明显的效果, 且普遍强于理想摘要对冗余度的要求. 然而, 过于追求低冗余会影响摘要的其他特性, 例如 DIVERSUM, FACES, FACES-E 和 LinkSUM 对冗余度的约束会导致无法选满 k 个三元组, 从而对摘要整体的重要性等性质产生影响. ESSTER 采用惩罚项的方式降低冗余, 一定程度上缓解了上述问题.

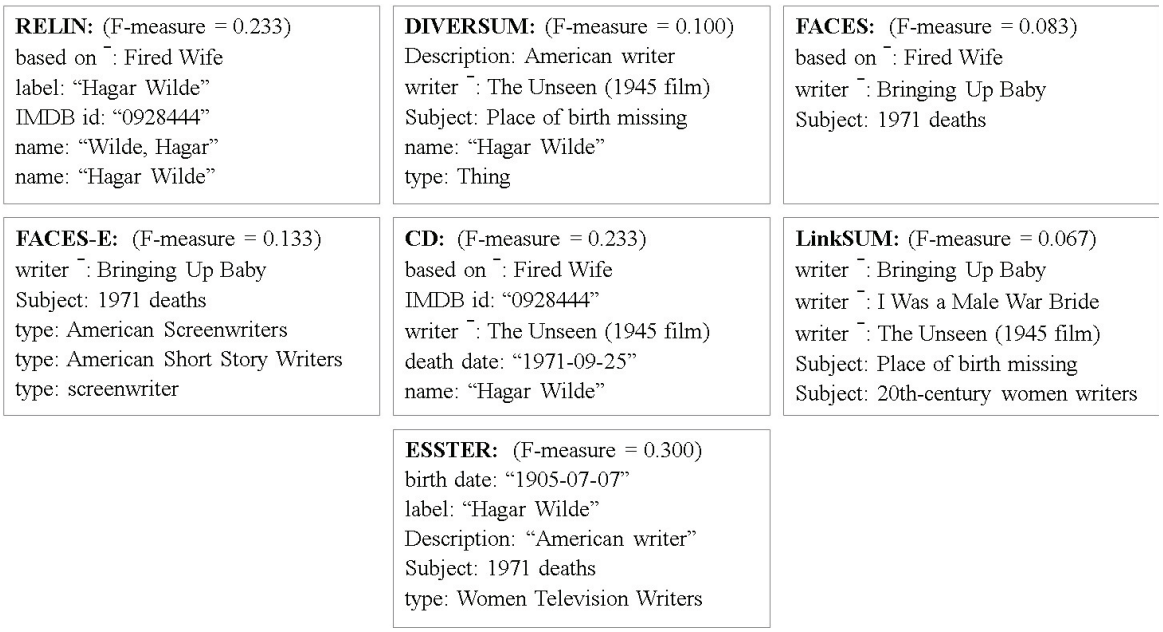

图 3 各方法为实体 `Hagar Wilde` 生成的摘要以及相应的 F-measure 得分

Figure 3 Summaries generated by entity summarizers for entity `Hagar Wilde`, along with their F-measure scores.

4.3 实例分析

本小节通过分析实例来比较各摘要方法的特点. 图 3 呈现了各摘要方法为实体 `Hagar Wilde` 生成的摘要, 并给出了相应的 F-measure 得分, 各三元组都用属性 – 值对的形式表示. 其中, 短线表示对应谓词的反向含义, 即 $\mathrm{pred}(t)^-$.

从实例结果看来, RELIN 由于缺乏对冗余的考虑, 生成的摘要中有 3 个三元组 (对应属性 `label` 和 `name`) 之间内容重复. 又由于其将自信息作为结构重要性而缺乏对可读性的考虑, 因而选择了 `IMDB id` 这种取值较为独特但可读性较低的属性. DIVERSUM 通过强约束选择具有不同属性的三元组, 但缺乏对结构重要性的考虑, 以至于选取了 `type:Thing` 这种在知识图谱中过于普遍存在而缺乏区分度的三元组, 相比于 ESSTER 选取的 `type:Women Television Writers`, DIVERSUM 的选择提供的信息有限. FACES 仅处理取值为实体的三元组, 并且要求同一聚类中属性相同的三元组不可同时选入, 这种约束导致其摘要只选了 3 个三元组. 其扩展方法 FACES-E 增加了对属性为 `TYPE` 和取值为字面量的三元组的处理, 但其基于层次聚类的方式可能让同属性的三元组属于不同聚类, 因此最终生成的摘要中出现了 3 个属性同为 `TYPE` 且相互之间含义存在一定重复的三元组. CD 考虑了去冗余, 所生成的摘要内没有重复内容, 但由于未考虑可读性, 选入了属性为 `IMDB id` 的三元组. LinkSUM 避免选择取值相同的三元组, 而未考虑属性间的冗余, 最终摘要中只包含了来自 `writer` 和 `Subject` 这两种属性的信息, 摘要内容的多样性有限. 相比之下, ESSTER 生成的摘要无冗余, 内容多样, 且兼顾了结构重要性和可读性, 取得的 F-measure 得分远高于其他方法所生成的摘要.

5 相关工作

数据摘要任务通常分为抽取式和抽象式两类. 前者从待摘要目标中选取内容来构成摘要, 而后者

生成的摘要则是对待摘要内容的概括或改写. 现有实体摘要研究主要采取抽取式, 即从实体描述中选取有限个三元组来构成摘要, 本文同样关注抽取式摘要.

5.1 实体摘要

实体摘要的研究近年来形成了较多研究成果, 涉及多种研究任务和应用场景. 包括特定领域的实体摘要, 如电影领域 [15, 16] 和时间轴 [17] 等, 其研究对象为特定类型的知识图谱实体. 面向特定任务的实体摘要, 如用于共指消解的实体摘要方法 C3D+P [5], 用于实体链接场景的 COMB [4] 等, 其生成的摘要特定于要辅助用户完成的下游任务. 上下文相关的实体摘要, 如以查询 [3] 或文档内容 [18] 为上下文, 摘要内容的选取需要考虑与上下文的相关度. 而更多的实体摘要研究则关注于通用场景, 即根据实体和知识图谱自身的内容和特性, 自动生成可广泛应用于多种领域和应用的实体摘要. 本文同样关注的是通用场景下的实体摘要, 并将主要讨论这一类相关方法.

Cheng 等 [6] 提出的 RELIN 方法综合考虑三元组的自信息和三元组之间的相关性, 构造带权 PageRank 模型对三元组打分排序. DIVERSUM [7] 倾向于选择带有高频属性的三元组, 并提出生成具备多样性的摘要, 要求选入摘要的三元组不可包含重复的属性. FACES [8] 基于词袋模型的相似度对三元组聚类, 并根据三元组的自信息和取值的流行度进行打分排序, 最终从尽可能多的聚类中选取打分高的三元组以构成包含多样内容的摘要. 其扩展方法 FACES-E [9] 在此基础上增加了对取值为字面量的三元组的处理. SUMMARUM [19] 和 LinkSUM [11] 关注取值为实体的三元组, 并根据取值的 PageRank 打分对三元组排序, LinkSUM 对三元组的打分还考虑了取值与目标实体间的 backlink, 并对属性排序以避免选取具有相同取值的三元组. CD [10] 将实体摘要建模为二次背包问题求解, 以提高所选三元组的统计信息量同时降低三元组间的冗余度为优化目标.

可以看出, 现有实体摘要方法主要由对三元组的重要性评价策略和对摘要的降冗余策略所构成. 在三元组打分策略上, DIVERSUM 主要考虑属性的频度, RELIN 和 CD 均考虑三元组的自信息, FACES 和 FACES-E 在此基础上考虑了取值的流行度, 而 LinkSUM 则对取值和属性分别打分. 本文提出基于图谱结构的三元组打分方法, 采用与以往不同的方式综合考虑了属性和取值在知识图谱全局和局部图结构中的表现, 以兼顾摘要内容的通用性与独特性. 在降冗余策略上, DIVERSUM 采用基于属性的强约束, LinkSUM 则是避免取值的冗余, 都可能导致摘要内容不足 k 个. FACES 和 FACES-E 采用相对弱化的约束, 通过从不同聚类中选取三元组以降低冗余. 本文提出的方法对三元组两两之间冗余度的计算考虑了文本、数值和逻辑推理 3 个方面, 并量化为优化目标中的惩罚项, 以适应不同容量条件下对冗余的不同容忍程度. 这一做法与 CD 对冗余的处理方式相近. 然而相比于 CD, 本文采用了不同的三元组重要性评价策略, 并且不同于以往所有方法, 本文考虑了对三元组可读性的计算.

5.2 其他相关摘要

数据摘要工作最早可追溯到对文档数据的摘要 [20]. 抽取式文档摘要任务从一个或多个文档中选取句子构成摘要. 传统文档摘要方法采用启发式或基于相似度对句子排序 [21]. 近年随着深度学习在文本处理任务中取得突出成果, 文档摘要领域的大量工作也采用了相关技术, 将对句子的选择建模为序列标注问题, 结合 seq2seq 等神经网络模型求解 [22]. 不同于文档摘要面对序列型文本, 实体摘要处理的是来自知识图谱的结构化数据. 三元组之间无相对顺序, 单个三元组相对文档句子所含文本更短, 且三元组中各元在图谱中有明确角色. 相比于文档摘要方法关注于挖掘文本序列中的语义, 实体摘要方法则可以利用三元组各元在知识图谱结构中的特征. 此外, 两者在生成摘要的目标上也有所不同. 通用场景的文档摘要需要涵盖目标文档所含主题的核心内容, 而实体摘要则会关注使其区别于其他实

体的独特内容. 尽管文档摘要与实体摘要有诸多不同, 其研究中的一些想法仍给予了本文工作一定的启发, 例如降低摘要的冗余 [23], 以及提高摘要的可读性 [24], 而本文给出了不同的实现方式.

实体摘要可视为一种特殊的图摘要. 一般的图摘要任务旨在对整个图规模的缩减 [25], 从而得到对整个图代表性内容的简洁表示. 因此, 抽取式的图摘要方法通过选取最有代表性的节点或边来构造摘要. 而非抽取式方法则生成对原图的高层抽象. 相比之下, 本文所关注的实体摘要问题其特殊之处在于, 待摘要的范围为图中一个特定节点 (即目标实体) 及其邻居导出的子图, 而非对整个图做摘要. 因此, 对方法技术的需求完全不同.

6 总结与展望

本文提出了通用场景下实体摘要问题的新方法 ESSTER. 为利用三元组内容在知识图谱中的结构特点, 提出基于属性和取值在全局和局部图数据中的流行度度量三元组的结构重要性. 为减少摘要中的冗余现象, 从逻辑冗余、数值冗余和文本冗余 3 个方面度量三元组之间的冗余程度. 此外, 考虑到普通用户对不同三元组内容的不同理解程度, 本文提出借助文本语料度量三元组可读性的方法. 最后, 综合这 3 个方面的度量, 将实体摘要任务转化为二次背包问题建模求解. 本文在两个公开数据集上将该方法与 6 种现有实体摘要方法进行了对比, 实验表明 ESSTER 所考虑的 3 部分度量均能为摘要质量带来显著提升, 其综合方法在各配置下取得与对比方法显著更优或相当的结果.

通用场景下的摘要任务的主要目标是满足大部分用户的需求. 而在未来的研究中, 需要考虑满足不同用户的个性化需求. 本文提出的方法为适应不同需求提供了可扩展的框架, 例如, 可通过修改 profit 中对各项的权重, 模拟不同用户的不同偏好. 在一些特定场景中, 可能由于面向的任务和领域的不同, 而存在对摘要的特殊要求, 相应就需要研究在重要性等方面结合其他因素, 例如用户查询, 应用上下文等等. 此外, 为了提高摘要在实际应用中的灵活性和适用性, 还需要考虑引入交互机制来获取用户的反馈, 并结合反馈深入理解用户的需求. 这就需要深入研究可利用反馈信息动态更新的摘要模型, 并设计合理的交互方式, 建模交互流程. 另外, 由于缺乏大规模监督数据, 现有实体摘要工作较少采用如深度神经网络等高复杂度模型. 为此, 实体摘要未来的工作还可考虑构建大规模标注数据集以支持复杂模型的学习, 或研究针对实体摘要场景的半监督或迁移学习等方式以充分利用有限的标注数据构造模型.

参考文献

1 Zhang L, Zhang Y, Chen Y. Summarizing highly structured documents for effective search interaction. In: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2012. 145–154

2 Yan J, Wang Y, Gao M, et al. Context-aware entity summarization. In: Proceedings of the 17th Internal Conference on Web-Age Information Management, Part I. Switzerland: Springer, 2016. 517–529

3 Hasibi F, Balog K, Bratsberg S E. Dynamic factual summaries for entity cards. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2017. 773–782

4 Cheng G, Xu D, Qu Y. Summarizing entity descriptions for effective and efficient human-centered entity linking. In: Proceedings of the 24th International Conference on World Wide Web. New York: ACM, 2015. 184–194

5 Cheng G, Xu D, Qu Y. C3D+P: a summarization method for interactive entity resolution. J Web Semant, 2015, 35: 203–213

6 Cheng G, Tran T, Qu Y. RELIN: relatedness and informativeness-based centrality for entity summarization. In: Proceedings of the 10th International Semantic Web Conference, Part I. Berlin: Springer, 2011. 114–129

7 Sydow M, Pikuła M, Schenkel R. The notion of diversity in graphical entity summarisation on semantic knowledge graphs. J Intell Inf Syst, 2013, 41: 109–149
8 Gunaratna K, Thirunarayan K, Sheth A P. FACES: diversity-aware entity summarization using incremental hierarchical conceptual clustering. In: Proceedings of the 29th AAAI Conference on Artificial Intelligence. California: AAAI Press, 2015. 116–122
9 Gunaratna K, Thirunarayan K, Sheth A P, et al. Gleaning types for literals in RDF triples with application to entity summarization. In: Proceedings of the 13th European Semantic Web Conference. Switzerland: Springer, 2016. 85–100
10 Xu D, Zheng L, Qu Y. CD at ENSEC 2016: generating characteristic and diverse entity summaries. In: Proceedings of the 2nd International Workshop on Summarizing and Presenting Entities and Ontologies. Ruzica Piskac: ESUR-WS.org, 2016
11 Thalhammer A, Lasierra N, Rettinger A. LinkSUM: using link analysis to summarize entity data. In: Proceedings of the 16th International Conference on Web Engineering. Switzerland: Springer, 2016. 244–261
12 Stoilos G, Stamou G B, Kollias S D. A string metric for ontology alignment. In: Proceedings of the 4th International Semantic Web Conference. Berlin: Springer, 2005. 624–637
13 Pisinger D. The quadratic knapsack problem—a survey. Discrete Appl Math, 2007, 155: 623–648
14 Yang Z, Wang G, Chu F. An effective GRASP and tabu search for the 0-1 quadratic knapsack problem. Comput Oper Res, 2013, 40: 1176–1185
15 Thalhammer A, Toma I, Roa-Valverde A J, et al. Leveraging usage data for linked data movie entity summarization. 2012. ArXiv: 1204.2718
16 Li Y, Zhao L. A common property and special property entity summarization approach based on statistical distribution. In: Proceedings of the 2nd International Workshop on Summarizing and Presenting Entities and Ontologies. Ruzica Piskac: ESUR-WS.org, 2016
17 Gottschalk S, Demidova E. EventKG — the hub of event knowledge on the web — and biographical timeline generation. Semant Web, 2019, 10: 1039–1070
18 Tonon A, Catasta M, Prokofyev R, et al. Contextualized ranking of entity types based on knowledge graphs. J Web Semant, 2016, 37–38: 170–183
19 Thalhammer A, Rettinger A. Browsing DBpedia entities with summaries. In: Proceedings of the 11th European Semantic Web Conference. Switzerland: Springer, 2014. 511–515
20 Jones K S. Automatic summarising: the state of the art. Inf Process Manag, 2007, 43: 1449–1481
21 Mihalcea R, Tarau P. TextRank: bringing order into text. In: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2004. 404–411
22 Zhou Q, Yang N, Wei F, et al. Neural document summarization by jointly learning to score and select sentences. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2018. 654–663
23 Carbonell J G, Goldstein J. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 1998. 335–336
24 Ganesan K, Zhai C, Viegas E. Micropinion generation: an unsupervised approach to generating ultra-concise summaries of opinions. In: Proceedings of the 21st World Wide Web Conference. New York: ACM, 2012. 869–878
25 Liu Y, Safavi T, Dighe A, et al. Graph summarization methods and applications: a survey. ACM Comput Surv, 2018, 51: 1–34

Entity summarization with high readability and low redundancy

Qingxia LIU, Gong CHENG* & Yuzhong QU

National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023*, China*
* Corresponding author. E-mail: gcheng@nju.edu.cn

Abstract The development of the World Wide Web has triggered substantial growth of knowledge graphs (KG). Research into using KGs for intelligent applications has increased significantly. A KG describes facts about entities using RDF triples, and an entity description may contain a large number of triples. In applications where entity information is presented directly, entity summarization is required to prevent user information overload and to fit the presentation capacity. Here, the task is to select the most representative subset of triples from the rich entity description. In this paper, we propose an innovative entity summarization method, which we refer to as ESSTER, to generate summaries with both high readability and low redundancy. The proposed method combines structural and textual features. The importance of a triple is measured based on its structural features in the KG. The text readability of a triple is measured based on n-grams in a text corpus, and redundancy in a set of triples is measured by logical reasoning, numeric comparison, and text similarity. Entity summations is modeled and by combining these three measures and solved as a combinatorial optimization problem. We conducted experiments and compared the proposed method to six existing methods on two publicly available datasets of manually labeled summaries. Experimental results demonstrate that the proposed method achieves state of the art results.

Keywords knowledge graph, entity summarization, redundancy, readability, combinatorial optimization

Qingxia LIU was born in 1990. She received a B.S. degree in computer science and technology from Nanchang University, Nanchang, in 2011. Currently, she is a Ph.D. student at the Department of Computer Science and Technology at Nanjing University. Her research interests include data summarization and question answering.

Gong CHENG was born in 1984. He received a Ph.D. degree in computer software and theory from Southeast University, Nanjing, in 2010. Currently, he is an associate professor at the Department of Computer Science and Technology at Nanjing University. His research interests include semantic search, data summarization, and question answering.

Yuzhong QU was born in 1965. He received a Ph.D. degree in computer software from Nanjing University, Nanjing, in 1995. Currently, he is a professor at the Department of Computer Science and Technology at Nanjing University. His research interests include semantic Web, question answering, and novel software technology for the Web.