噪声超链接移除系统的实现：语义与相关性方法

Kazem Taghandiki*、Elnaz Rezaei Ehsan†

*计算机工程系，技术与职业大学（TVU），伊朗德黑兰 ktaghandiki@tvu.ac.ir

† 硕士，工业工程、系统管理与生产力，伊朗科技大学 elnazrezaeie110@gmail.com

arXiv:2303.03321v1 [cs.IR] 6 Mar 2023

摘要——随着万维网上的数据量不断增长，作为万维网图形化表示的网络结构图也在持续演变。该图的结构已逐渐从基于内容转向非基于内容。此外，网络结构图中的垃圾数据（如噪声超链接）会对信息检索和链接挖掘算法的速度与效率产生不利影响。以往该领域的研究主要采用结构方法和字符串方法来移除噪声超链接。然而，在某些情况下，这些方法可能会误删有用链接，或无法检测出噪声超链接。本文首先使用交互式爬虫构建超链接数据集，随后借助语义网方法和工具（如 DBpedia本体）研究超链接的语义与相关性结构。最后，使用 DBpedia本体上的推理器执行噪声超链接移除过程。实验结果证明，语义网技术能够准确且有效地移除噪声超链接。

Index 术语—语义网、噪声超链接、本体、推理器、语义相似性、关联相似性 .

1 引言

近年来，除了在各个技术领域产生的海量数据之外，创建任意数量网页的能力 [1], 催生了一个颇具挑战性的概念，即“大数据” [2], [3]。2021 年，Google 和必应网络爬虫索引了近 490 亿个网页 [4]。显然，互联网中的网页数量显著激增，进而推动了网络结构图的增长。因此，由于噪声超链接等垃圾数据的存在，浏览和探索万维网的结构变得极其困难 [1]。因此，显然需要一种消除垃圾超链接的机制。历史上，许多信息检索算法利用网络文档的内容对垃圾网页进行分类、聚类和删除。然而，这种处理密集且耗时的方法后来被依赖网络结构图中超链接特征而非文档内容的算法所取代。PageRank 等多种算法 [5] 利用这些特征来减少搜索引擎所需的处理量。然而，这些基于链接的算法以链接准确指向用户所需页面这一假设为前提 [6], [7]。因此，许多链接挖掘算法错误地假设网络结构图完全基于语义和内容。然而，该图中包含无用的垃圾链接，这些链接既可能误导用户，也可能影响算法的输出。垃圾超链接会使不相关的文档获得比相关文档更高的排名。这种提高页面排名的尝试主要是出于“商业”目的 [6]。已有多项研究致力于从万维网结构中检测和消除垃圾链接。然而，所提出的方法高度依赖于超链接的字符串和结构特征 [6], [8] ，而忽略了其语义和关联结构。本文从页面和站点两个层级考察超链接的语义和关联结构。利用语义网技术，例如本体和推理器，我们删除噪声超链接。为此，首先通过一个独立的过程创建超链接数据集。随后，利用本体和推理器等语义网技术，对源页面超链接的概念以及目标页面的概念进行语义和关系分析。该分析可用于判定某个超链接是噪声超链接还是有用超链接。

所提出的系统以构建的数据集作为输入；数据集的每一行由以下内容组成：从源页面的超链接上下文主题映射而来的类别、从目标页面主题映射而来的类别、一个表示从用户视角来看该超链接属于噪声超链接还是有用超链接的字段，以及源页面的域名。随后，系统会检测噪声超链接，并将结果与用户的判断进行比较，以确定本体中的每个语义属性或关联属性在正确检测中所发挥的作用。此外，该系统还具有其他优势，例如可以了解哪些查询会将用户引导至噪声超链接，以及哪些域名包含数量最多的噪声超链接

[9]。实验表明，语义网技术在消除噪声超链接方面具有准确率、能力和可扩展性。本文其余部分的组织结构如下。第2节综述了以往关于移除噪声超链接的研究。第3节详细介绍了所提出方法的实现。第4节阐述了实验及所得结果。最后，在第5节中给出结论。

2 相关工作

关于超链接移除系统的实现，有几种可行的研究类型，例如，[10],[11]。在这方面，Qi 等人 [12]将导航、广告和无关超链接归为垃圾信息，而其他超链接则被视为有用。在他们的算法中，使用具有两个类（即“合格”和“不合格”）的支持向量机（SVN），通过总计六种字符串相似度特征来检测和过滤噪声超链接。将该算法应用于包含 210 万个网页的集合后，23% 的超链接被归类为“不合格”。不过，该机制并未采用语义或相关性方法来移除超链接。Wookey 等人[13] 设计了一个名为 Anchor Woman 的系统，该系统利用网站的超链接结构检测噪声超链接，并将其分为三类：

- 多弧环：在网络图中形成多个循环的超链接链。
    - 多弧：许多指向同一页面的超链接。
    - 递归环：包含指向自身的超链接的网页。

该系统接收一个网址作为输入，并对其超链接进行广度优先搜索，以识别并移除噪声超链接。随后，系统会为用户生成网络结构的图形化层次表示。Carvalho 等人提出了一种在站点层级检测噪声超链接的机制。 [14] 他们识别出两类垃圾关系：（1）相互强化，即两个网站通过交换站点层级的超链接而紧密相连；（2）由一系列紧密相连的网站构成的联盟。如果网站之间的超链接数量超过阈值，就会被视为噪声。该算法采用结构化方法，能够移除 16.7% 的超链接，平均精度均值为 59.16。Chakrabarti [15]提出了一种更精细的万维网模型，其中页面以其文档对象模型表示，生成的 DOM 树再由常规超链接相互连接。该方法能够抵御“裙带关系团体攻击”，但与我们的算法（完全基于链接分析）相比，需要更多输入数据。此外，由于我们专门针对噪声移除，因此能够识别不同类型的超链接异常。Samanta 等人 [?] 采用基于图的方法改进网络结构图并促进用户导航。该论文研究了英国大学网站的情况。研究人员从 110 个学术网站中提取了 600 多万个链接，以构建一个数据集。借助 TextPipe，他们移除了大量指向图像、音频和视频文件的非预期链接。网络文档数量、路径长度和强连通分量（SCC）均通过这种方法进行了优化。然而，所提出的方法仅依赖超链接类型，并未考虑语义或关系方法。这两个步骤如下：

1. 删除通常位于页面顶部附近的广告和导航超链接。

2. 删除不属于关联关系或聚合关系的超链接。

聚合关系是指两个概念之间的层级关系，其中源概念的范围比目标概念更广。关联关系也称为水平关系，意味着源概念和目标概念具有相同的父节点。换言之，当且仅当两个概念具有共同的父节点时，它们才具有水平关系。作者报告称，移除导航超链接的识别率为 92.89%。聚合关系体现了一种语义方法；然而，关联关系不能被视为完整的相关性方法。尽管 Pedersen 等人 [16]使用关联关系或水平关系来表明两个概念之间的关联相似性，但水平关系仅表示两者之间的“部分-整体”关系。尽管如此，本体中的其他关系属性（例如对象属性）也可用于表示关联相似性。Oguz [17] 于 2022 年提出了另一种用于移除噪声超链接的算法，称为网站结构提取算法（WSE）。WSE 的主要目标是在保留语义超链接的同时移除噪声超链接。然而，该论文在语义方面关注的是超链接的路径结构，以维持超链接的层级关系。例如，假设有四个页面 A、B、C 和 D，那么从 A 指向 B、从 B 指向 C，以及从 C 指向 D 的超链接属于语义超链接，而方向相反的超链接则被视为噪声超链接。在 [18] 中，作者提出使用语言模型检测裙带关系链接。在该方法中，如果根据源页面和目标页面的语言模型判断二者不相关，就会降低该链接的权重。这种方法基于以下假设：由非裙带关系链接连接的页面必须具有足够的相似性。Wu 和 Davison [19]提出了一种两步算法来识别链接农场。第一步根据网页入链与出链的交集生成种子集。第二步扩展种子集，将指向种子集中多个页面的页面纳入其中。随后，对这些已识别垃圾网页之间的链接重新加权，并对修改后的链接图应用排名算法。以往研究往往聚焦于页面级超链接；然而，现代垃圾网站通常利用站点级超链接，通过生成指向其他网站的非法链接来提升其在 Google 索引中的排名。因此，考虑站点级超链接至关重要。本文旨在同时移除页面级和站点级的噪声超链接。以往研究的另一项不足在于只采用字符串方法或结构方法。尽管这些方法检测超链接类型的速度相当快，但有时会移除有用的超链接，有时又无法检测到噪声超链接。例如，假设一个包含文本“Bank”（意为金融机构）的页面指向一个介绍“Banks”（意为河岸）的页面。使用字符串

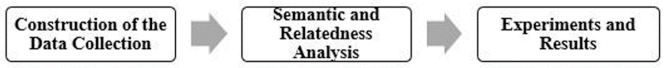

图 1. 构建数据收集

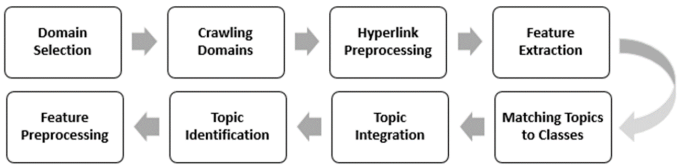

图 2. 数据收集的构建阶段

方法时，该超链接会被视为有用；而语义方法则利用超链接文本信息来识别并移除噪声超链接。本文旨在应用语义网方法和 DBpedia 本体等现有工具，考察超链接的语义与关系结构，并通过激活 DBpedia 本体推理机来移除噪声超链接。实验表明，语义网技术能够有效且准确地移除噪声超链接。在前述机制以及其他机制中，例如 [1], [20], [21], [22], [23], [24],，现有数据集并未用于信息检索；相反，噪声超链接是通过分析网络结构图来移除的。因此，我们需要构建一个超链接数据集。下一节将根据信息检索原则 [25] 介绍这一流程的详细内容。

3 提出的方法

用于消除噪声超链接的语义与相关性系统包括三个主要步骤，如图 1 所示。

3.1 构建数据收集

数据集构建是一个独立的过程，由几个步骤组成，如图 2 所示。

在构建数据集时，用户仅参与域选择，其他步骤均独立完成。

3.1.1 领域选择

互联网用户每天出于各种目的使用网络搜索引擎，向 Google 等搜索引擎提交数百万条查询。在此，决定抓取哪些网站是一个重要问题，因为噪声超链接和有用超链接都是必要的。前者将利用本体的语义和相关性属性予以消除，而后者则有助于证明所提出方法利用相同属性检测有用超链接的能力。所抓取的网站必须包含在互联网用户中广受欢迎的内容。利用 Google Trends，识别出了多个热门主题，包括新闻、金钱、在线购物、新技术，以及演员或运动员等名人。根据 Google Trends，表1展示了2021年的热门搜索查询。

表 1 2021 年热门搜索查询

News	名人	体育	电子产品
阿富汗	詹妮弗·劳伦斯	皇家马德里足球俱乐部	iPhone 13
AMC股票	金·卡戴珊	切尔西足球俱乐部	Galaxy Z Flip4
新冠疫苗	朱莉·盖耶	巴黎圣日耳曼足球俱乐部	Nexus Summit
狗狗币	Tracy Morgan	巴塞罗那足球俱乐部	摩托罗拉 Moto G Power



图 3. 网络爬虫提取的链接

通过了解互联网用户中的热门主题，组织和个人可以在设计网页时采取两种不同的方法。

1. 创建主题相关性较弱的网站，却利用后台超链接开展利润丰厚的商业活动，例如将用户引导至在线商店或色情网站。这类网站包含与万维网用户需求毫不相关的噪声超链接。

2. 创建包含特定主题有用内容的网站，为用户提供恰当的信息。这些网站中的超链接很少被视为垃圾信息。它们包含符合用户需求的有用超链接。这种域名选择方法的核心思路是爬取搜索引擎针对高频查询返回的域名。返回的域名大致分为两类：（1）包含以非法商业目的设置的噪声超链接的域名；（2）以合法目的运营、提供有用超链接来帮助用户实现目标的域名。区分这两类域名的能力体现了所提方法的优势：在移除噪声超链接的同时保留有用超链接。

3.1.2 域名爬取

在此阶段，用户将上一阶段的每个主题输入 Google 搜索引擎，并从返回的网站中随机选择一部分。然后，将网站地址作为输入，交给一个使用 Java 编程语言库开发的交互式爬虫。接下来，爬虫开始探索该域；最后，便可获得该域中的链接列表，如图 3 所示。

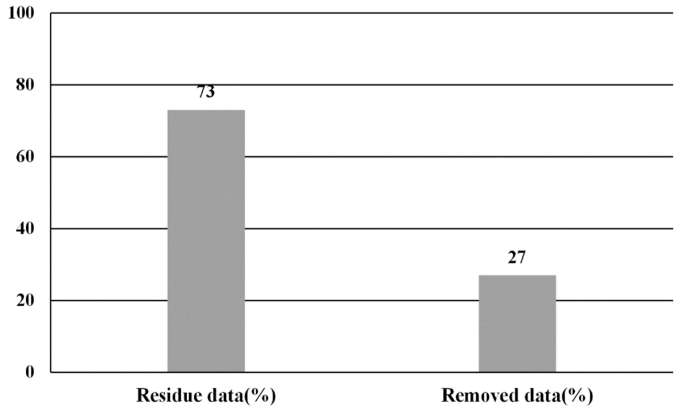

图4. 预处理后超链接的状态

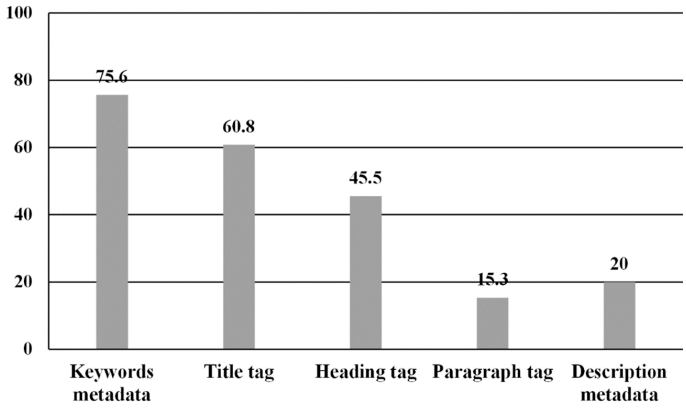

图 5. 网页中最常用的五种特征

在此处，网络爬虫共爬取了 114 个与这些主题相关的有用和无用域名，这些域名均为英文。

3.1.3 超链接预处理

提取出的超链接中有很大一部分不适用于本文的目的，例如重复链接或指向音频、视频文件的链接。因此，根据 [1]（第2节）提出的操作将其移除。预处理后超链接的状态如图4所示。

如图所示，经过预处理操作后，27%的超链接被移除，爬取的超链接数量从2665个减少到1946个。

请注意，提取出的域名中有很多包含指向视频、图像和音频的超链接，这些链接与所提出的方法不兼容。在某些情况下，同一页面中会多次出现相同链接。因此，27%的链接（719个链接）被删除。如前所述，所提出的方法仅适用于文本超链接，因此后续步骤（即与数据集构建相关的步骤）也仅适用于文本超链接。因此，需要在预处理步骤中移除不兼容的超链接。

3.1.4 特征提取

为了确保能够以足够的准确率检测出超链接周围文本的主题以及其目标页面的主题，必须从网页中提取若干特征。然而，我们需要确定设计网页时使用频率较高的特征，这有助于识别需要提取的关键特征。图 5 展示了在 5000 个页面中五种关键特征的出现频率。如图所示，“关键词元数据”、“标题标签”和“一级标题标签”是网页中最常见的特征。

因此，使用页面的三个特征，即标题标签、关键词元数据和一级标题标签，来确定目标页面的主题。除了前两个特征外，还利用超链接文本及其所在段落，从源页面中提取超链接的主题上下文。超链接文本及其周围的段落构成了超链接上下文的特征。表2和表3分别展示了用于确定目标页面主题和超链接文本上下文的特征示例

表 2 检测目标页面主题所提取的特征（来自 www.Filehippo.ocm）

页面标题	免费下载软件
关键词元数据	下载软件 免费软件 共享软件程序
一级标题	软件

用于分别确定目标页面主题和超链接文本上下文的提取特征。

这些特征加快了后续阶段的分析和主题检测流程。与我们的工作不同，若干研究 [26], [27], [28], [29], [30], [31] 基于文档的全部内容提取主题。大多数网页设计师利用 HTML 等 Web 2.0 技术来创建网络文档。与 Web 3.0 的 RDF 、XML 和 OWL 等技术相比，这种标记语言在语义表达能力方面较弱 [29]。因此，鉴于 HTML 属性在设计网络文档时广受欢迎，这种语言似乎是 Web 2.0 中进行特征选择的最佳选择。然而，借助 Web 3.0 技术，语义化提取特征变得相当容易，而这对所提出的方法发挥着重要作用。不过，这超出了本文的讨论范围，仍是未来研究值得探索的方向。

3.1.5 特征预处理

必须提高上一步提取的特征质量，以便语义与相关性系统能够以更高的准确率和更低的错误率执行主题识别过程。本文通过采用典型的文本挖掘预处理技术来实现这一目标，例如停用词处理、词元规范化、大小写折叠和词干提取 [25]:

1. 停用词是指频繁出现且对于查找特定信息作用不大的词。示例包括冠词和介词，如“the”“and”“or”、

表 3 检测超链接文本主题上下文所提取的特征（来自 www.Filehippo.com）

页面标题	免费下载软件
关键词元数据	下载软件 免费软件 共享软件 程序
超链接文本	新软件
超链接段落	新软件的最新版本

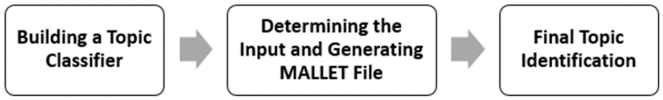

图 6. 有监督的主题检测过程

等。移除这些词可加快主题检测步骤。

2. 词元规范化是一种标准化过程，旨在为每个单词采用统一的形式。例如，“anti discriminatory”和“nondiscriminatory”经过规范化处理后，都会映射为“anti-discriminatory”。

3. 大小写折叠是一种广为人知的单词规范化过程，其中大写字母会转换为对应的小写字母。事实上，大小写折叠可视为词元规范化的一种。

4. 根据语法规则，单词通常会采用不同形式；例如，organize、organize 和 organize。此外，一个单词的不同形式可能具有近似的含义，例如 democracy 和 democratic。词干提取过程通过去除单词的词尾，旨在为单词的不同形式提取共同词根。本研究使用 MALLET 以及 Python 中的文本挖掘库进行预处理。

3.1.6 主题识别

本阶段旨在利用先前提取的特征，确定超链接文本和目标页面的主题。如图 6 所示，本阶段通过包含三个操作的有监督过程完成。所有操作均使用 MALLET 执行。以下小节将介绍相关细节。

1. 构建主题分类器

为了构建主题分类器，需要一组高质量的初始特征。因此，我们决定从 5000 个网页中提取三种最常见的关键特征（即关键词元数据、标题标签和一级标题标签），以此构建特征数据集。主题分类器基于该数据集构建。此外，80% 的数据用于训练，其余 20% 用于测试分类器的准确率。本文采用四种不同的方法进行主题分类，即朴素贝叶斯、C4.5、决策树和最大熵，并进行 10 次交叉验证 [32], [33], 以获得较高的效率。我们对这些算法进行了比较，以找出最佳分类方法。根据图 7，最大熵在训练数据上取得了最高准确率。因此，我们使用该算法来构建主题分类器。

2. 确定输入并生成MALLET文件 此操作将从超链接文本上下文和目标页面中提取的特征作为输入提供给 MALLET，以识别各自的主题。此操作通过命令行界面完成。由此可获得两个输出文件，也称为特征向量。这些向量是对输入值的数值表示，可加快分析操作[34]。这些文件将作为下一项操作的输入。

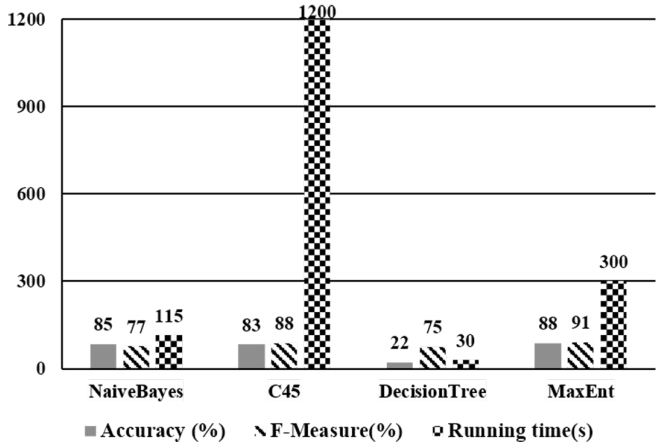

图 7. 朴素贝叶斯、决策树、C4.5 和最大熵在特征数据集上的准确率与执行时间性能比较

表4 最终主题识别步骤后获得的示例输出

文件4	文件3	文件2	文件1
National Geographic	0	Fish	Bird
National Geographic	0	哺乳动物	Bird
乳制品食品	1	画布	Car
MP3歌曲	1	电影	Song
MP3歌曲	1	Game	音乐

3. 最终主题识别

在此步骤中，获取上一项操作生成的 MALLET 输出文件，并由主题分类器识别主题。此外，后续步骤可以使用生成的模型（即输出）推断新输入数据的主题。完成后，主题分类器能够成功识别约 81.8% 的已提取特征的主题。其余 18.2% 的特征可能因以下原因之一而无法分类：

1. 由于设计不佳，特征提取步骤可能无法提取适当的特征，以识别超链接文本和目标页面的主题。这种情况下将无法分配主题。

2. 用于构建主题分类器的特征可能与从新页面提取的特征完全不同。因此，该页面不会被分配主题。此步骤的输出包括四个独立的文本文件。表 4 展示了每个文件的一部分。

3.1.7 主题整合

将前一阶段位于多个文本文件中的主题合并整合，生成一个格式良好的单一输入，供匹配阶段使用。输出文件示例如图 8 所示。

从左到右，文件中的字段分别表示源页面中超链接的上下文主题、目标页面的主题、用户感知的链接类型（即噪声链接或有用链接），以及源页面所属的域。实际上，此阶段构建了一个主题数据集。

```
Bird,Fish,0,nationalgeographic
Bird,Mammal,0,nationalgeographic
Car,Canvas,1,dairyfoods
Song,Movie,1,songsmp3
Music,Game,1,songsmp3
Wine,Food,0,dairyfoods
```

图 8. 主题集成输出文件

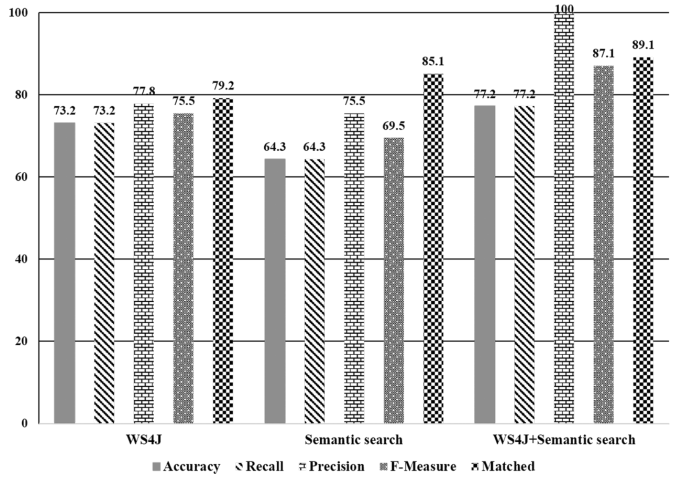

图 9. WS4J、SemanticSearch 与 WS4J+语义搜索的比较

3.1.8 将主题匹配至类

此阶段的目的是将上一阶段得到的主题与 DBpedia本体中的类进行匹配。匹配操作使用 WS4J 库和术语搜索算法（SemanticSearch）[35]完成。图 9 展示了使用 400 个随机选取的主题和五项标准（包括准确率）对这两种算法进行的比较。

根据结果，WS4J 与语义搜索可能具有互补性，因此有必要考察二者的同时应用。如图 9 所示，WS4J 与语义搜索的组合优于所有单一算法。此外，这两种算法都支持所需的语义相似度和字符串相似度。因此，我们使用 WS4J+语义搜索进行匹配。最终匹配结果的示例见图 10。

http://dbpedia.org/ontology/Automobile	http://dbpedia.org/ontology/Automobile	0	ebay
http://dbpedia.org/ontology/Currency	http://dbpedia.org/ontology/Actor	1	mydailyfundose
http://dbpedia.org/ontology/Continent	http://dbpedia.org/ontology/Actor	1	mydailyfundose
http://dbpedia.org/ontology/Activity	http://dbpedia.org/ontology/Actor	1	mydailyfundose
http://dbpedia.org/ontology/Coach	http://dbpedia.org/ontology/Actor	1	mydailyfundose
http://dbpedia.org/ontology/MusicFestival	http://dbpedia.org/ontology/Actor	1	mydailyfundose
http://dbpedia.org/ontology/Economist	http://dbpedia.org/ontology/Actor	1	mydailyfundose
http://dbpedia.org/ontology/Food	http://dbpedia.org/ontology/Actor	1	mydailyfundose
http://dbpedia.org/ontology/NaturalPlace	http://dbpedia.org/ontology/Actor	1	mydailyfundose

图 10. 匹配示例

再次说明，这些字段分别表示与源页面中超链接的主题上下文相匹配的类别、目标页面中匹配类别的主题、超链接类型以及域名。在此步骤中，使用映射子系统，近 87.3% 的主题被映射到 DBpedia 本体中的相应主题。事实上，在此步骤中创建了一个最终数据集，可称为概念超链接数据集。

3.2 语义与相关性分析

在此阶段，将最终的数据收集结果作为系统的输入，其中包含两个概念或本体类：超链接文本上下文和目标页面。随后，系统利用从 DBpedia本体中添加的知识，检验输入内容的关联性和语义属性。通过这种方式，可以判断超链接是噪声链接还是有用链接。推理器是语义与相关性分析步骤中最关键的组件。它是一种软件，作用于使用本体创建的一个或多个概念数据集。推理器旨在从本体中现有的事实提取逻辑结果。所提出的系统使用 Pellet 完成这一任务。Pellet 推理器基于 DBpedia本体中的基本知识，通过本体的属性、关系和类获取新增知识。接下来，将数据收集结果中的一行（代表网页中的一个超链接）以两个概念的形式输入，即超链接文本上下文和目标页面。利用从 DBpedia本体推断出的属性和关系，确定二者的语义相似性和相关性相似性。在本文其余部分中，我们分别使用“源概念”和“目标概念”来指代超链接文本和目标页面的概念。

如果至少满足以下属性之一，推理器就会将超链接视为有用超链接：

1. “等价类”：源概念与目标概念等价。2. “子类”和“具有超类”：源概念是目标概念的子类或超类。

属性（1）和（2）称为语义属性，用于表示两个概念之间的语义相似性。例如，“女性”和“人”这两个概念在语义上相似，因为前者是后者的子类。

3. “对象属性”：源概念和目标概念通过对象属性相关联。这称为关联属性，表示关联相似性。例如，“猴子”和“香蕉”这两个概念通过“喜爱”或“食用”等多种关系具有关联相似性（即“猴子喜欢香蕉”或“猴子吃香蕉”）。目前许多噪声超链接移除方法只关注前两种属性，而忽略对象属性。因此，这些方法无法检测源概念和目标概念之间的关联属性，进而可能错误地移除有用超链接。

如果这些条件都不成立，那么源概念与目标概念之间就不存在语义相似性或相关性相似性，因此该超链接被视为噪声。换句话说，该超链接没有将用户带到其想访问的页面，而是从当前页面跳转到一个意料之外且完全无关的页面。因此，在这一步结束时，一个

http://dbpedia.org/ontology/Automobile	http://dbpedia.org/ontology/Automobile	subClassOf	0	ebay
http://dbpedia.org/ontology/Cartoon	http://dbpedia.org/ontology/Agent	animator	0	wikipedia
http://dbpedia.org/ontology/Person	http://dbpedia.org/ontology/Place	birthPlace	0	wikipedia
http://dbpedia.org/ontology/Person	http://dbpedia.org/ontology/Place	livingPlace	0	wikipedia
http://dbpedia.org/ontology/Film	http://schema.org/Movie	equivalentClass	0	athlete
http://dbpedia.org/ontology/MusicFestival	http://dbpedia.org/ontology/Actor		1	mydailyfundose
http://dbpedia.org/ontology/Economist	http://dbpedia.org/ontology/Actor		1	mydailyfundose
http://dbpedia.org/ontology/Food	http://dbpedia.org/ontology/Actor		1	mydailyfundose

图 11. 分析阶段构建的记录

表 5 语义与相关性系统的混淆矩阵

推断标签				
实际标签 由用户提供		Yes	NO	总计
	Yes	1145（真正例）	232（假负例）	1377
	NO	50（假正例）	519（真负例）	569
	总计	1195	751	1946

超链接概念数据集便构建完成，其噪声或有用属性由推理器确定。图 11 展示了分析阶段构建的若干条记录。

从左到右，各字段分别表示主语、宾语、前两个字段之间关系的推断关系属性和语义属性、推理器感知到的超链接类型，以及源页面的域。

3.3 实验与结果

混淆矩阵是评估所提出方法的一种重要且实用的工具，其中涉及两种标注：（1）语义与相关性分析期间的系统标注；（2）创建数据集时的专家用户标注。矩阵中的每个元素可能属于以下类别之一：真正例（TP）、真负例（TN）、假正例（FP）和假负例（FN）。表 5 展示了系统的混淆矩阵，该矩阵基于数据收集阶段收集的用户意见，以及推理器在语义与相关性分析期间得出的推断生成。

- TP 表示用户和所提出的系统均认为有用的超链接数量。
- FP 表示用户认为有噪声、而所提出的系统认为有用的超链接数量。
- TN 表示用户和所提出的系统均认为有噪声的超链接数量。
- FN 表示用户认为有用、而所提出的系统认为有噪声的超链接数量。图 12 展示了信息检索系统中六项常用性能指标的数值。如图所示，所提出的方法准确率和精确率较高，同时将错误率维持在足够低的水平。

图 13 展示了推理器使用 DBpedia本体中的各种语义和相关性属性，在主语与宾语之间表示语义与相关性相似性的程度。

从图 13 可以看出，在 12.6% 的情况下，推理器使用“子类”属性将源超链接概念与目标页面概念关联起来。“等价类”对应的比例为 36.7%。这些属性表明源概念与目标概念之间存在语义相似性。此外，在 11.9% 的情况下，推理器使用“对象属性”，这

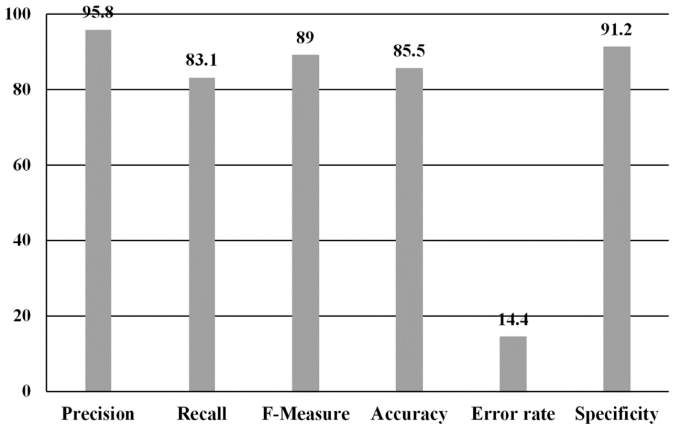

图 12. 所提出方法的性能指标

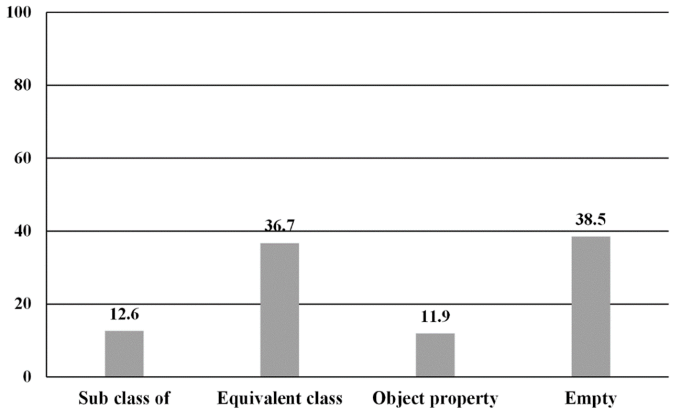

图 13. 推理器所使用的 DBpedia本体属性百分比

表示关联相似性。对于其余 38.5% 的情况，使用 DBpedia 本体未能发现语义属性或关联属性。表 6 展示了通过语义与相关性分析获得的结果概览。

特征提取步骤会获取并存储与超链接域（¡a¿ 标签的 href 值）相关的一小组信息。因此，用户在源页面上遇到超链接后，会判断该超链接是否为噪声。我们的目标是将结果与用户的看法进行比较。为更好地理解表 6 中的语义和关系分析结果，还将其与

表 6 语义与相关性分析结果

爬取链接数量	数据集中的链接数量
2665	1946
域名数量	本体类
114	312（38%）
有用域名	
维基百科、BBC、Facebook、YouTube、eBay	
噪声域名	
mydailyfundose, Google, fullmoviesfreedownload, hollywoodlife, 维基百科	
有用域名 噪声域名	
61.4%	38.55%
噪声链接的目标概念	
购物中心、货币、电影、演员、药物	

表7 从用户视角分析数据收集

爬取链接数量	数据集中的链接数量
2665	1946
域名数量	
114	
有用域名	
维基百科、BBC、Facebook、YouTube、eBay	
噪声域名	
mydailyfundose、Google、fullmoviesfreedownload、hollywoodlife、songmp3	
有用域名	噪声域名
70.75	29.2%
噪声链接的目标概念	
购物中心、货币、电影、演员、模特	

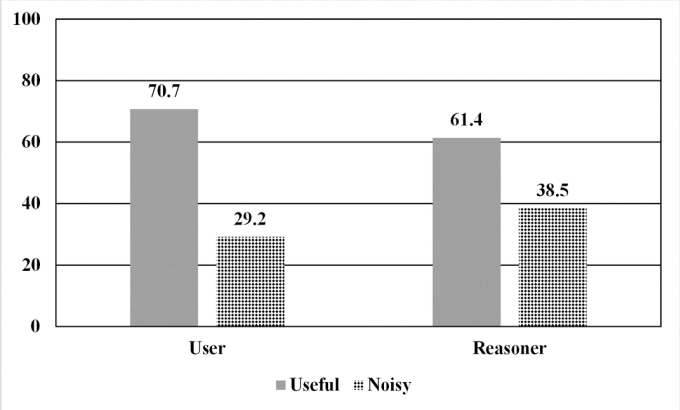

图14. 用户与推理器所感知的超链接类型百分比

用户视角进行比较。表7概述了用户视角，用于判断超链接是否有用或属于噪声。

对表6和表7的比较证明了语义与相关性方法的能力。图14展示了从这两个角度观察到的超链接类型百分比。从图中可以明显看出，所提出的方法能够像专家用户一样准确地区分噪声超链接和有用超链接。

图14显示，推理器和用户分别在61.4%和70.7%的情况下检测出了有用超链接。对于噪声超链接，这一比例分别为38.5%和29.2%。根据这些结果可以明确看出，推理器能够达到专家用户的水平。

3.3.1 可扩展性

最终的数据集必须包含能够通过语义与相关性方法进行分析的记录，这一点至关重要。换言之，数据集必须包含属于DBpedia本体中各类的记录。推理器需要一定时间，根据DBpedia本体发现新的关系、类和属性，并确定数据集中超链接的类型（即噪声超链接或有用超链接）。本节余下部分将讨论推理器以及整个流程的可扩展性。可扩展性测试在一台计算机上进行，该计算机配备 Intel Core i5 2450M（ 2.50 GHz）处理器、4.00 GB内存，并运行64位操作系统。图15展示了针对500、1000、1500和1946个超链接运行语义与相关性推理器所需的时间。

变量 y 和 R 分别表示超链接的数量和方程的准确率。在

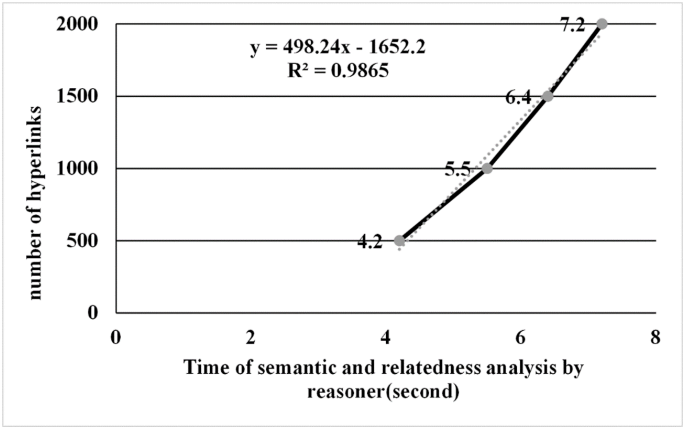

图 15. 不同数量超链接下推理器的可扩展性

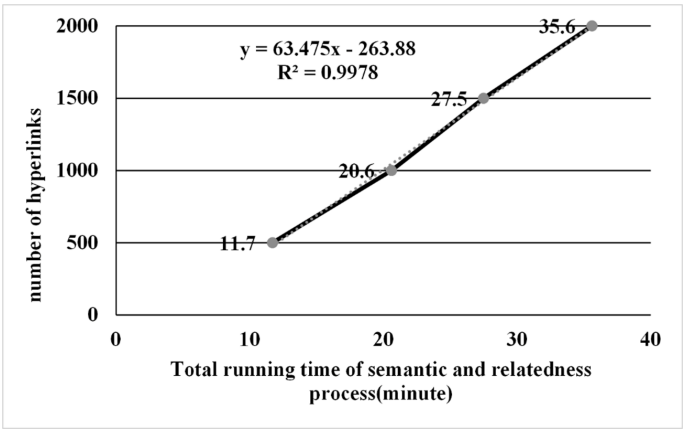

图 16. 不同数量超链接下整个系统的可扩展性

最初的两到三秒内，推理器正在启动。图 15 假设最终数据集已可用。然而，图 16 针对相同的情况，考察了从主题识别到语义与相关性分析整个过程的可扩展性。

我们的研究发现，匹配阶段是整个过程中最耗时的步骤。

4 讨论与结论

超链接是万维网结构中的一种噪声数据，会对许多信息检索算法的效率产生负面影响。几乎所有这类算法都关注超链接的字符串或图结构。因此，这些方法会错误地删除某些有用超链接，并且在某些情况下无法检测出噪声超链接。本文在页面和站点层级上考察语义结构与关系结构，同时运用本体和推理器等语义网技术来消除噪声超链接。为此，通过一个独立流程创建超链接数据集。随后，从语义和关联性两个方面对该数据集进行分析。分析结果可区分噪声超链接和有用超链接。所提出的系统以构建的数据集为输入。数据集的每一行均包含：从超链接映射得到的类

源页面的上下文主题所映射到的类、从目标页面主题映射得到的类、一个用于标明从用户视角看该超链接是噪声还是有用的字段，以及源页面的域名。随后，将结果与用户的判断进行比较，以展示本体中的各项语义属性或关系属性在将超链接识别为噪声或有用超链接方面所起的作用。此外，还可以找出会引导用户访问噪声超链接的查询类别，以及噪声超链接数量最多的域名。我们的实验验证了语义网技术在消除噪声超链接方面的准确率、能力和可扩展性。未来工作方向包括：1. 将 DBpedia本体与另一种本体结合，以在 T-Box 层级涵盖更广泛的域概念。2. 使用现有关联数据集来涵盖 A-Box 层级的概念。3. 将噪声检测操作扩展至指向图像、视频和音频文件的超链接。4. 应用各种算法，将主题与本体类进行匹配。5. 使用语义工具，借助采用 Web 3.0 技术构建的页面中的语义属性，识别超链接和目标页面的主题。

参考文献

[1] D. Samanta、S. Dutta、M. G. Galety 和 S. Pramanik，《一种面向高性能计算的网络挖掘分类体系新方法》，载于 网络智能与信息检索：*CIIR 2021* 论文集。施普林格，2022 年，第 425–432 页。[2] M. Sisk、M. Majlis、C. Page 和 A. Yazdinejad，《分析可解释人工智能指标：文献综述总结》，2022 年。[3] A. Yazdinejad、A. Dehghantanha、R. M. Parizi、G. Srivastava 和 H. Karimipour，《安全智能模糊区块链框架：物联网网络中的有效威胁检测》，《工业计算机》，第 144 卷，第 103801 页，2023 年。[4] M. de Kunder，《Worldwidewebsize.com——万维网（互联网）的规模》，*worldwidewebsize*，2015 年。[5] Y. Wu、Y. Wu、Y. Liu 和 T. Shi，《基于加权 PageRank 算法的 Raft 共识算法优化方案研究》，载于 *2022* 年亚洲算法、计算与机器学习会议（*CACML*）。电气电子工程师学会，2022 年，第 784–789 页。[6] S. Solanki、S. Verma 和 K. Chahar，《PageRank 算法综合研究》，载于 计算智能演进：第九届智能计算前沿：理论与应用国际会议（*FICTA 2021*）论文集。施普林格，2022 年，第 1–10 页。[7] A. Yazdinejad、R. M. Parizi、G. Srivastava 和 A. Dehghantanha，《理解人工智能深度伪造技术中的区块链》，载于 *2020* 年电气电子工程师学会 *Globecom* 研讨会（*GC Wkshps*。电气电子工程师学会，2020 年，第 1–6 页。[8] E. Elakkiya 和 S. Selvakumar，《用于社交网络垃圾信息检测（son2s）的前馈神经网络分层超参数优化》，《软计算》，第 26 卷，第 21 期，第 11 915–11 934 页，2022 年。[9] A. Yazdinejad、A. Dehghantanha、R. M. Parizi 和 G. Epiphaniou，《一种基于 NSGA-II 的数据分类优化模糊深度学习模型》，《神经计算》，第 522 卷，第 116–128 页，2023 年。[10] S. Nakhodchi、B. Zolfaghari、A. Yazdinejad 和 A. Dehghantanha，《Steeleye：工业控制系统中利用半深度学习进行应用层攻击检测与归因的模型》，载于 *2021* 年第 *18* 届隐私、安全与信任国际会议（*PST*）。电气电子工程师学会，2021 年，第 1–8 页。[11] A. Yazdinejad、A. Bohlooli 和 K. Jamshidi，《Virtex-6 FPGA ML605 上 OpenFlow v1.3 交换机的高效设计与硬件实现》，《超级计算杂志》，第 74 卷，第 1299–1320 页，2018 年。[12] X. Qi、L. Nie 和 B. D. Davison，《通过测量相似性检测合格链接》，载于 第三届网络对抗性信息检索国际研讨会论文集，2007 年，第 49–56 页。

[13] L. Wookey 和 J. Geller，《面向网页搜索者的网站结构语义层次抽象》，《信息技术研究与实践期刊》，第 36 卷，第 1 期，第 23–34 页，2004 年。[14] A. L. da Costa Carvalho、P.-A. Chirita、E. S. De Moura、P. Calado 和 W. Nejdl，《面向搜索引擎的站点层级噪声去除》，载于第 15 届万维网国际会议论文集，2006 年，第 73–82 页。[15] S. Chakrabarti，《将文档对象模型与超链接相结合，以增强主题提炼和信息提取》，载于第 10 届万维网国际会议论文集，2001 年，第 211–220 页。[16] T. Pedersen、S. Patwardhan、J. Michelizzi 等，《WordNet：相似性——概念相关性的度量》，载于 AAAI，第 4 卷，2004 年，第 25–29 页。[17] R. F. Oguz、M. Oz、E. Olmezogullari 和 M. S. Aktas，《从大规模图数据中提取信息：自动化用户界面测试案例研究》，载于《Euro-Par 2021：并行处理研讨会：Euro-Par 2021 国际研讨会，葡萄牙里斯本，2021 年 8 月 30–31 日，修订精选论文》。施普林格，2022 年，第 364–375 页。[18] A. A. Bencz´ur、I. B´ır´o、K. Csalog´any 和 M. Uher，《通过语言模型分歧检测裙带关系链接》，载于第 15 届万维网国际会议论文集，2006 年，第 939–940 页。[19] B. Wu 和 B. D. Davison，《识别链接农场垃圾网页》，载于第 14 届万维网国际会议专题轨道与海报，2005 年，第 820–829 页。[20] A. Yazdinejad、B. Zolfaghari、A. Dehghantanha、H. Karimipour、G. Srivastava 和 R. M. Parizi，《工业物联网边缘设备中的精准威胁狩猎》，《数字通信与网络》，2022 年。[21] S. Bechhofer、F. van Hermelen、J. Hendler、I. Horrocks、D. McGuinness 和 P. Patel-Schneider，《W3C：OWL Web本体语言——参考》，2004 年。[22] K. Viswanathan 和 A. Yazdinejad，《虚拟现实系统的安全注意事项》，arXiv 预印本 arXiv:2201.02563，2022 年。[23] M. Keller 和 M. Nussbaumer，《超越网络图：利用导航结构图挖掘万维网的信息架构》，载于 2011 年新兴智能数据与网络技术国际会议。电气电子工程师学会，2011 年，第 99–106 页。[24] A. Yazdinejad、A. Dehghantanha、R. M. Parizi、M. Hammoudeh、H. Karimipour 和 G. Srivastava，《区块猎手：面向基于区块链的工业物联网网络的网络威胁狩猎联邦学习》，IEEE工业信息学汇刊，第 18 卷，第 11 期，第 8356–8366 页，2022 年。[25] C. D. Manning，《P. Raghavan 和 H. Schütze，《信息检索导论》》，2008 年。[26] A. Yazdinejad、R. M. Parizi、A. Dehghantanha 和 K.-K. R. Choo，《P4到区块链：面向软件定义网络的安全的区块链赋能数据包解析器》，《计算机与安全》，第 88 卷，第 101629 页，2020 年。[27] G. Ercan 和 I. Cicekli，《使用词汇链进行关键词提取》，《信息处理与管理》，第 43 卷，第 6 期，第 1705–1714 页，2007 年。[28] J. Kupiec 、 J. Pedersen 和 F. Chen，《可训练的文档摘要器》，载于第 18 届 ACM SIGIR 信息检索研究与发展年度国际会议论文集，1995 年，第 68–73 页。[29] D. Sheridan、J. Harris、F. Wear、J. Cowell Jr、E. Wong 和 A. Yazdinejad，《Web3 市场面临的挑战与机遇》，arXiv 预印本 arXiv:2209.02446，2022 年。[30] B. Lott，《关键词提取技术综述》，UNM Education，第 50 卷，第 10 期，2012 年。[31] S. Robertson，《理解逆文档频率：关于 IDF 理论论证的探讨》，《文献学杂志》，2004 年。[32] A. Yazdinejad、R. M. Parizi、A. Dehghantanha 和 H. Karimipour，《用于无人机身份认证的联邦学习》，《自组织网络》，第 120 卷，第 102574 页，2021 年。[33] Z. Nematzadeh、R. Ibrahim 和 A. Selamat，《使用机器学习技术和 K 折交叉验证进行乳腺癌分类的比较研究》，载于 2015 年第十届亚洲控制会议（ASCC）。电气电子工程师学会，2015 年，第 1–6 页。[34] A. Yazdinejad、R. M. Parizi、A. Bohlooli、A. Dehghantanha 和 K.-K. R. Choo，《使用 P4 和 FPGA 的网络可编程数据包处理器高性能框架》，《网络与计算机应用杂志》，第 156 卷，第 102564 页，2020 年。

[35] B. Zolfaghari, A. Yazdinejad 和 A. Dehghantanha，《云与物联网的二分法：从安全视角看云辅助物联网》，*arXiv* 预印本 *arXiv:2207.01590*，2022。