Postprint of Research on Entity Importance Ranking in News Documents

Authors: Lu Na, Zhou Pengcheng, Wuchuan

Date: 2023-08-26T00:00:00+00:00

Abstract

[Purpose/Significance] Existing research on entity ranking in news documents predominantly adopts a document-centric or entity-centric perspective, focusing on tasks such as text classification and entity linking, while studies explicitly addressing entity importance within texts remain relatively scarce. This research investigates importance-based entity ranking for news documents. [Method/Process] Given a document, we determine the importance of entities relative to the document and rank them accordingly. Experiments are conducted on the Sogou comprehensive web news dataset, with entity ranking results evaluated using NDCG and inverse order comparison ratio as metrics. [Results/Conclusion] Experimental results demonstrate that methods based on entity frequency, TF*IDF, information entropy, TextRank, and ensemble methods all achieve favorable performance, whereas the method based on clustering coefficient yields only moderate effectiveness. Specifically, the TF*IDF-based method achieves an NDCG value of 95.86%, representing the best performance under this metric; the ensemble method achieves an inverse order comparison ratio of 84.46%, representing the best performance under this metric.

Full Text

Preamble

Research on Entity Importance Ranking for News Documents

Lu Na[1], Zhou Pengcheng[2], Wu Chuan[2]
[1]School of Information Science and Technology, Hainan Normal University, Haikou 571158
[2]School of Information Management, Wuhan University, Wuhan 430072

Abstract

[**Purpose/Significance**] Existing research on entity ranking in news documents primarily focuses on document-centric or entity-centric tasks such as text classification and entity linking, with relatively few studies addressing entity importance within texts. This study investigates importance-based entity ranking for news documents. [**Method/Process**] Given a document, we assess the relative importance of entities within that document and rank them accordingly. Experiments are conducted on the Sogou comprehensive news dataset, with evaluation performed using NDCG and inverse pair rate metrics. [**Results/Conclusions**] Experimental results demonstrate that methods based on entity frequency, TF-IDF, information entropy, TextRank, and ensemble approaches all achieve favorable performance, while the clustering coefficient method yields moderate results. The TF-IDF method achieves an NDCG value of 95.86%, representing the best result for that metric; the ensemble method achieves an inverse pair rate of 84.46%, representing the best result for that metric.

Keywords: news documents, entity importance, entity ranking

Introduction

Entities represent a special semantic unit in documents that contain rich semantic information and have received increasing attention in recent years. Current entity-related research includes named entity recognition [**?**], entity linking [**?**], and entity relation extraction [**?**], all of which fall under the umbrella of information extraction. Named entity recognition identifies text fragments representing named entities, which include seven categories: person names, location names, organization names, percentages, times, dates, and currencies. Entity linking connects text fragments representing entities to specific entries in knowledge bases such as Wikipedia, Freebase, and YAGO. Typically, named entity recognition serves as the first step in entity linking—determining the boundaries of entity text fragments before using entity disambiguation methods to uniquely identify entities and link them to a given knowledge base. Entity relation extraction extracts structured data from unstructured text, represented as subject-predicate-object triples in the form <Entity1, Relation, Entity2>, where Entity1 and Entity2 are key entity concepts. However, specialized analysis of entity importance within documents remains underexplored.

This study focuses on entity importance within documents: given a document, we determine the relative importance of its constituent entities and rank them accordingly. Traditional entity ranking primarily falls into two categories: related entity ranking and query-oriented entity ranking. Related entity ranking involves finding entities in the entire entity set that satisfy certain conditions relative to a given entity. Query-oriented entity ranking returns entities most relevant to a web query. Both approaches operate across all entities in a document collection. In contrast, our task ranks entities within a single document

based on their importance to that document—a fundamentally different problem.

News documents represent one of the most common text types on the Internet and contain a relatively large number and variety of entities compared to other document types. Consequently, this study selects news documents as its research context. Since entity-document importance relationships have received limited attention, to ensure feasibility and comprehensibility, we limit entity types to three categories: person, location, and organization.

The concept of entity importance has received some attention in existing research. For example, M. Liu et al. [**?**] mentioned entity importance in their news summarization work, identifying important sentences and key entities based on importance. Like our study, M. Liu et al. focused on entity importance within single documents. However, their research did not provide a concrete definition of key entities and emphasized the news summarization problem itself, utilizing multiple information sources including query terms and news titles. In contrast, our study focuses specifically on entity importance. M. Liu et al. [**?**] proposed a method for identifying important events and key entities within a time window, judging importance based on novelty (an entity's rising trend in that window) and popularity (its frequency). Their approach emphasized macro-level analysis of entity importance in the overall environment, whereas our study focuses on judging entity importance within single documents, independent of time and trends.

This research holds theoretical significance: by analyzing entity importance in documents, it helps text analysis tasks identify focus areas (important entities) and reduce noise (peripheral entity interference), thereby facilitating entity-oriented knowledge organization and preventing irrelevant documents from being considered when mining entity information. Practically, this research can be applied to news classification/clustering and news recommendation. By determining entity importance and incorporating entity attributes from knowledge bases, we can assist various text mining tasks. In portal websites, if user interest in specific entities is detected, we can prioritize recommending documents where those entities are important, thereby increasing click-through rates.

Literature Review

"Time," "location," and "person" constitute the fundamental elements of news reporting. Some Chinese scholars have explored the importance of news elements in news summarization research. Guo Yanqing et al. [**?**] proposed a news summarization method based on event element weighting, treating time, location, person, and organization as news event elements (what we call entities). They used the ICTCLAS segmentation system from the Chinese Academy of Sciences to identify event elements and weighted these elements by frequency across multiple news documents. Wu Lingda et al. [**?**] utilized basic local topic sentence groups and extended local topic sentence groups in multi-document

summarization, where basic local topic sentence groups were generated by first assigning weights to basic news elements (time, location, person, organization) using TF-IDF values, then applying clustering methods. These studies share some similarities with our work in examining entity importance in news documents. However, our study applies different methods to a given single document to determine the relative importance of different entities and rank them accordingly.

Foreign researchers have introduced the concept of key entities for news summarization and other tasks. For example, K. Kiritoshi et al. [**?**] defined key entities in news documents as the most important set of entities and identified them using TF-IDF. However, their research focused on news recommendation—ranking other news documents by relevance to a given news document—whereas our study ranks entities within a single document by importance. M. Liu et al. [**?**] studied entity-based news summarization, constructing a graph with query terms, news titles, sentences, and entities as nodes, defining four relationships (title-sentence, query-sentence, sentence-sentence, sentence-entity), and applying PageRank to judge sentence and entity importance. Their approach utilized multiple information sources, while our study focuses specifically on entity importance itself.

Methods for Importance-Based Entity Ranking

Entity importance ranking for news documents involves extracting entities from a news document, assessing their importance using specific methods, and ranking them accordingly. Formally, given input document d containing entities $\{e_1, e_2, ..., e_n\}$ where e_i represents the i-th entity, the output is an entity list $e(1) > e(2) > ... > e(n)$ where $e(i)$ is the entity ranked i-th in importance. Assuming news document $d = \{p_1, p_2, ..., p_n\}$ where p_i is the i-th paragraph, we first preprocess d by segmenting paragraphs into sentences and performing named entity recognition. Metrics for judging entity importance include entity frequency, TF-IDF, clustering coefficient, information entropy, and TextRank. This study proposes methods based on these individual metrics as well as an ensemble method that weights and averages three local feature indicators (entity frequency, distribution entropy, and entity TextRank values in co-occurrence networks) multiplied by the global feature IDF.

Entity Frequency Method

A fundamental hypothesis about entity importance in documents states: if entity e_i appears more frequently in news document d, it is likely more relevant to d and thus more important. Since we examine entity importance within single documents, local features are particularly crucial. We count entity occurrences and normalize by total entity occurrences in the document, as shown in Formula (1):

$$EF_i = \frac{count_i}{\sum_{j=1}^{n} count_j} \tag{1}$$

where $count_i$ represents the frequency of entity e_i in the document. EF (entity frequency) measures entity importance relative to other entities in the document.

The entity frequency method is simple, intuitive, and interpretable. However, its limitation becomes apparent when multiple entities share the same frequency—without additional information, ranking them is difficult. This limitation is more pronounced in shorter news texts.

TF-IDF Method

Inverse document frequency (IDF) measures an entity's discriminative capability. The TF-IDF method combines entity frequency and inverse document frequency to compute importance, as shown in Formula (2):

$$TF\text{-}IDF_i = EF_i \cdot \log \frac{N}{DF_i + 1} \tag{2}$$

where EF_i represents entity frequency, N is the total number of documents in the collection, and DF_i is the number of documents containing entity e_i. To avoid division by zero, we apply add-one smoothing by adding 1 to all entity document frequencies.

While EF is a single-document local feature, IDF is a global feature. Combining local and global features comprehensively captures entity importance from both perspectives.

Clustering Coefficient Method

D. Beeferman [**?**] and T. R. Niesler [**?**] discovered word clustering phenomena in information retrieval and part-of-speech recognition research, where word distances follow exponential decay and words more relevant to documents exhibit more significant clustering. Similarly, entities of different importance levels should display different clustering characteristics, enabling us to use clustering indicators to judge importance.

We calculate entity clustering coefficients using spatial distribution in documents. Specifically, assuming entity e_i appears at positions $\{p_0, p_1, ..., p_{n-1}, p_n, p_{end}\}$ in document d (where e_i appears n times, p_i denotes the starting position of the i-th occurrence with $p_0 = 0$ representing document start, and p_{end} represents document end), the distances between occurrences are $\{p_2 - p_1, p_3 - p_2, ..., p_n - p_{n-1}\}$. The average distance is:

$$l = \frac{\sum_{i=1}^{n-1}(p_{i+1} - p_i)}{n-1} = \frac{p_n - p_1}{n-1} \tag{3}$$

The standard deviation of distances is:

$$s = \sqrt{\frac{\sum_{i=1}^{n-1}((p_{i+1} - p_i) - l)^2}{n-2}} \tag{4}$$

The clustering coefficient is defined as:

$$c = \frac{s}{l} \tag{5}$$

where l is the average distance and s is the standard deviation. A larger clustering coefficient indicates greater entity importance.

This method assumes that important entities appear frequently in local document regions, resulting in small average distances. Therefore, we compute clustering coefficients for all entities in news documents and rank them in descending order. Notably, document length significantly affects clustering coefficients, so we normalize entity positions: if entity e_i's i-th occurrence is at position p_i in document d of length l, we normalize it to p_i/l, enabling calculation using normalized position information.

Information Entropy Method

In 1949, Shannon proposed the concept of information entropy. For all possible events from an information source with probabilities $\{p_1, p_2, ..., p_n\}$, Shannon proposed an uncertainty metric $s(p_1, p_2, ..., p_n)$:

$$s(p_1, p_2, ..., p_n) = -K \sum_{i=1}^{n} p_i \log p_i \tag{6}$$

where K is a constant. Information entropy has the following properties: (1) $s(p_1, p_2, ..., p_n)$ is continuous; (2) when $p_i = 1/n$, $s(p_1, p_2, ..., p_n)$ reaches its maximum and is monotonically increasing with n.

Assuming document d consists of n parts where part i contains n_i entities and entity e_i appears $n_i(e_i)$ times in part i, the total entity occurrences in d is $\sum_{i=1}^{n} n_i$, and entity e_i's relative frequency in part i is $f_i(e_i) = n_i(e_i)/n_i$. We define the probability distribution of entity e_i in part i as $p_i(e_i) = f_i(e_i)/\sum_{j=1}^{n} f_j(e_j)$. According to Shannon's theory, the information entropy of entity e_i's distribution is:

$$S(e_i) = -\frac{\sum_{i=1}^{n} p_i(e_i) \ln p_i(e_i)}{\ln(n)} \tag{7}$$

where $\ln(n)$ is the constant K ensuring $S(e_i)$ falls between 0 and 1. By dividing documents into paragraphs and calculating information entropy for all entities, we rank entities in descending order to achieve importance-based sorting.

TextRank Method

For input document $d = \{s_1, s_2, ..., s_m\}$ containing n entities $\{e_1, e_2, ..., e_n\}$, we first construct a weighted undirected graph $G(V, E)$ where V represents the node set (document entities e_i) and E represents the edge set determined by entity co-occurrence relationships. Specifically, if entities e_i and e_j co-occur in sentence S_k, their co-occurrence count increments by 1. We calculate edge weight $E(e_i, e_j)$ using Formula (8):

$$E(e_i, e_j) = \begin{cases} \frac{cooccur(e_i, e_j)}{\sum_{e_k \in Set_{e_i}} cooccur(e_i, e_k)} & \text{if } cooccur(e_i, e_j) > 0 \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

where $cooccur(e_i, e_j)$ denotes the co-occurrence count of entities e_i and e_j, and Set_{e_i} represents all entities co-occurring with e_i. This weighting scheme reflects that frequently co-occurring entities share specific semantic relationships, with higher co-occurrence indicating stronger semantic relatedness.

We then apply the PageRank algorithm to compute node PR values:

$$PR(e_i) = (1 - d) + d \cdot \sum_{e_k \in Set_{e_i}} W_{ik} \cdot PR(e_k) \tag{9}$$

where $PR(e_i)$ is the PageRank value of entity e_i, d is the damping coefficient (set to 0.85 following standard PageRank practice), and W_{ik} is the weight of edge $E(e_i, e_k)$. This formula embodies PageRank's core principle: entities co-occurring with more entities are more important, and entities co-occurring with important entities are themselves more important.

Ensemble Method

Entity importance may be determined by multiple factors, making single-metric assessments potentially biased. While entity frequency reflects importance, high frequency does not guarantee importance for every document. Additionally, some frequently occurring entities may appear only in document sections with minimal overall impact. Therefore, entity distribution across documents constitutes another crucial factor.

We propose an ensemble method that weights and averages three local feature indicators—entity frequency, information entropy, and entity TextRank values in co-occurrence networks—then multiplies by the global IDF feature. The ensemble method computes entity importance using Formula (10):

$$c\text{-}index = (a \cdot EF + b \cdot Entropy + c \cdot TR) \cdot IDF \tag{10}$$

where a, b, and c are weights for local features with $a + b + c = 1$, determined heuristically.

News Document Entity Importance Ranking Experiments

News Dataset

We evaluate our approach using the comprehensive news dataset from Sogou Labs [**?**]. This dataset collects news data from 18 channels (domestic, international, sports, society, entertainment, etc.) during June-July 2012, comprising over 1,290,000 news documents. We preprocess news content by segmenting sentences and performing named entity recognition, filtering out documents without textual content.

Annotation Dataset

Entity Importance Level Definition Different entities hold varying importance levels within documents, ranging from those most critical to document content to those mentioned only once and unrelated to the document theme. Following the classification in literature [**?**], we categorize entities into four importance levels: core entities (level 4), important entities (level 3), weakly relevant entities (level 2), and peripheral entities (level 1). The importance hierarchy is: core entities > important entities > weakly relevant entities > peripheral entities. Core entities are those around which the document revolves or whose relevance to the document is significantly higher than other entities. Important entities play substantial roles in news documents. Weakly relevant entities are not directly related to the document but connect directly to other document entities. Peripheral entities are simply mentioned with low news relevance.

Annotation Results We randomly selected 50 documents from the news corpus for manual entity importance level annotation, stored in a specific format shown in Figure 1 [Figure 1: see original paper]. Finally, we evaluate different entity importance ranking methods using NDCG and inverse pair rate metrics.

Evaluation Metrics

NDCG NDCG (Normalized Discounted Cumulative Gain) is a common information retrieval evaluation metric. It operates on two basic assumptions, adapted for our context as: (1) more important entities are more useful for

representing document main information; (2) lower-ranked entities have diminishing value as they are not core representatives of document information.

For an entity ranked at position n, NDCG is calculated as:

$$N(n) = Z_n \sum_{j=1}^{n} \frac{2^{r(j)} - 1}{\log(1+j)} \tag{11}$$

where Z_n is a normalization factor ensuring $N(n) \in [0, 1]$; $r(j)$ is the importance level of the j-th result; $2^{r(j)} - 1$ is the contribution value of the j-th result (importance levels and contribution values are shown in Table 1); and $\log(1+j)$ is the position discount (log base 2). For incorrectly recognized named entities, annotators assign importance level 0 with zero contribution value.

Inverse Pair Rate Assuming entities e_i and e_j in a news document where e_i's importance level exceeds e_j's, ideally e_i should rank before e_j, forming a correct order pair $< e_i, e_j >$. If e_i ranks after e_j, $< e_i, e_j >$ becomes an inverse order pair. The inverse pair rate is the proportion of inverse order pairs to correct order pairs in the annotated data. Entities with identical importance levels are neither correct nor inverse order pairs.

Relationship Between Metrics Both NDCG and inverse pair rate evaluate entity importance ranking, showing positive correlation but not linear relationship. For example, two methods might produce the same number of inverse order pairs (and thus identical inverse pair rates), but if one method inverts a core entity with an important entity while another inverts a weakly relevant entity with a peripheral entity, the former's NDCG will be lower due to greater discount on more important entity misplacements.

Experimental Results and Analysis

Table 2 presents entity importance ranking experimental results. The TF-IDF method achieves the highest NDCG value, differing from the entity frequency method by incorporating the global IDF feature. This improves performance, with NDCG increasing by 2.71% and inverse pair rate improving by 2.6%. The ensemble method incorporating global features also performs well, achieving an inverse pair rate of 0.8446—the highest among all methods.

Table 2. Entity importance ranking experimental results

Method	NDCG	Inverse Pair Rate
Entity Frequency	0.9333	0.8183
TF-IDF	0.9586	0.8396
Clustering Coefficient	0.7573	0.8015
Information Entropy	0.9269	0.8062

Method	NDCG	Inverse Pair Rate
TextRank	0.9158	0.8183
Ensemble Method	0.9578	0.8446

Results show the clustering coefficient method performs moderately, while other methods achieve good performance. The clustering coefficient method assumes larger coefficients indicate greater importance, but this assumption does not always hold. In some documents, peripheral entities exhibit high clustering coefficients while core entities show low coefficients.

Conclusion

Building upon existing research, this study conducts entity importance ranking experiments using entity frequency, TF-IDF, clustering coefficient, information entropy, TextRank, and ensemble methods. Compared to existing research, we address a relatively novel problem—single-document entity importance ranking—and provide preliminary exploration. We process the Sogou comprehensive news dataset, define entity importance levels, randomly sample 50 documents for annotation, and evaluate results using NDCG and inverse pair rate. Experimental results demonstrate that methods based on entity frequency, TF-IDF, information entropy, and TextRank achieve good performance, while the clustering coefficient method performs moderately.

Despite achieving certain results, this study has limitations. We only consider three entity types (person, location, organization), though some news categories (e.g., "health") may not contain these entities. The entity concept is broad, encompassing both concrete things (names, places, organizations) and abstract concepts (relations, ideas) whose attributes are also significant for news document processing. Future research should incorporate more entity types for greater 合理性 and significance. Additionally, our data relies on manual annotation, which may introduce subjective bias. Ideally, relevant labels should be mined from user-generated content to facilitate machine learning algorithm training and large-scale evaluation for improved effectiveness.

References

[1] Zhang Xiaoyan, Wang Ting, Chen Huowang. Research on named entity recognition [J]. Computer Science, 2005, 32(4): 44-48.

[2] Lu Wei, Wu Chuan. Survey on entity linking [J]. Journal of the China Society for Scientific and Technical Information, 2015(1): 105-112.

[3] Che Wanxiang, Liu Ting, Li Sheng. Automatic entity relation extraction [J]. Journal of Chinese Information Processing, 2005, 19(2): 1-6.

[4] Liu M, Liu Y, Xiang L, et al. Extracting key entities and significant events from online daily news [C]//Proceedings of the 9th international conference on

intelligent data engineering and automated learning. Berlin: Springer-Verlag, 2008: 201-209.

[5] Tran I S, Lucchese C, Perego R, et al. SEL: a unified algorithm for salient entity linking and saliency detection [C]//Proceedings of the 2016 ACM symposium on knowledge acquisition and modeling. New York: ACM, 2016: 85-94.

[6] Guo Yanqing, Zhao Rui, Kong Xiangwei, et al. News summarization extraction method based on event element weighting [J]. Computer Science, 2016(1): 237-241.

[7] Wu Lingda, Lei Zhen, Lao Songyang, et al. Event-related multi-document summarization based on local topic sentence groups [J]. Computer Simulation, 2006, 23(11): 263-267.

[8] Kiritoshi K, Ma Q. Named entity oriented related news ranking [C]//Proceedings of the 49th annual meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2011: 83-92.

[9] Liu M, Liu Y, Xiang L, et al. Single Chinese news article summarization based on entity information [C]//Proceedings of the second SIGIR workshop on Chinese language processing-volume 17. Stroudsburg: Association for Computational Linguistics, 2003: 184-187.

[10] Beeferman D, Berger A, Lafferty J. A model of lexical attraction and repulsion [C]//Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 1997: 373-380.

[11] Niesler T R, Woodland P C. Modelling word-pair relations in a category-based language model [C]//1997 IEEE international conference on Acoustics, Speech, and Signal Processing. Piscataway: IEEE, 1997, 2: 795-798.

[12] Sogou Labs. Comprehensive news data [EB/OL]. [2017-03-16]. http://download.labs.sogou.com/dl/.

[13] Pantel P, Fuxman A. Jigs and lures: associating web queries with structured entities [C]//Proceedings of the 49th annual meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2011: 84-88.

[14] Lin T, Etzioni O. Entity linking at web scale [C]//Proceedings of the joint workshop on automatic knowledge base construction and Web-scale knowledge extraction. Stroudsburg: Association for Computational Linguistics, 2012: 84-88.

[15] Welty C, Murdock J W, Kalyanpur A, et al. A comparison of hard filters and soft evidence for answer typing in Watson [C]//Proceedings of the 11th international conference on the Semantic Web-volume part II. Berlin: Springer-Verlag, 2012: 243-256.

[16] Zhang H P, Yu H K, Xiong D Y, et al. HHMM-based Chinese lexical analyzer ICTCLAS [C]//Proceedings of the second SIGHAN workshop on Chinese language processing-volume 17. Stroudsburg: Association for Computational Linguistics, 2003: 184-187.

[17] Shannon C E, Weaver W, Wiener N. The mathematical theory of communication [J]. Philosophical Review, 1949, 27(4): 623-656.

[18] Croft W B, Metzler D, Strohman T. Search engines: information retrieval in practice [M]. Beijing: China Machine Press, 2009: 1254-1271.

Author Contributions

Lu Na: Research design, experiments, manuscript drafting and revision
Zhou Pengcheng: Research design and revision, manuscript revision
Wu Chuan: Assisted in design, participated in manuscript revision

English Abstract

Importance-Based Entity Ranking for News Documents

Purpose/Significance: We propose an importance-based method for entity ranking. Entities in a particular document show different importance. Many researches focus on documents or entities, such as text categorization and entity linking, while few research pay attention to the importance of entities in documents. This research has significant theoretical and practical value. **Method/Process:** Given a document which consists of words and entities, our method computes the relative importance of entities in the document, and then ranks these entities based on their importance with respect to the document. We perform experiment on the Sogou News dataset, and use evaluation metrics such as NDCG and inverse pair rate to evaluate the results. **Result/Conclusion:** Experimental results show that methods based on entity frequency, TF-IDF, distribution entropy and TextRank achieve better performance, while method based on cluster coefficient does not work well. In terms of NDCG, TF-IDF method reaches 95.86%, which is the best result and in terms of the inverse rate, the ensemble method reaches 84.46%, which is the best result.

Keywords: news documents, entity importance, entity ranking

Note: Figure translations are in progress. See original paper for figures.