Research on Deep Learning-Based Valuation Methods for Cyber Science and Technology Intelligence Postprint

Authors: Zhang Min, Liu Huan, Ding Liangping, Fan Qing

Date: 2023-04-01T00:00:00+00:00

Abstract

[Purpose/Significance] Addressing the problem that researchers cannot timely identify intelligence content with intelligence value from massive online scientific and technological information, this paper establishes a comprehensive intelligence value calculation method to evaluate the intelligence value of online scientific and technological information, ultimately helping researchers quickly and accurately discover online scientific and technological information with intelligence value.

[Method/Process] Comprehensively considering both external features of intelligence and textual semantic content features, we utilize deep learning (pretrained language model) BERT method to construct an intelligence value calculation model based on textual semantic content features, use the prediction output of the deep learning model to complete scoring, and combine it with the original calculation method based on external features of intelligence to obtain the final comprehensive evaluation score.

[Results/Conclusion] Experimental results show that the intelligence value calculation model based on textual semantic content features can effectively classify intelligence into star ratings according to intelligence value scores, compensating for the poor star-rating discriminability in the original calculation model based on external features of intelligence. The final comprehensive evaluation results indicate that the intelligence value calculation model proposed in this paper can also well meet the needs of researchers in practical applications.

Full Text

Research on the Calculation Method of Web Technology Information Value Based on Deep Learning

Zhang Min[1,2,3,4], Liu Huan[2,3], Ding Liangping[2,3], Fan Qing[5]

[1]Wuhan Library, Chinese Academy of Sciences, Wuhan 430071
[2]National Science Library, Chinese Academy of Sciences, Beijing 100190
[3]Department of Library, Information and Archives Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190
[4]Hubei Key Laboratory of Big Data in Science and Technology, Wuhan 430071
[5]National Cultural Industry Research Center, Central China Normal University, Wuhan 430079

Abstract

[**Purpose/Significance**] To address the challenge that researchers face in timely identifying valuable intelligence content from massive amounts of web technology information, this paper establishes a comprehensive information value calculation method. This method evaluates the intelligence value of web technology information, ultimately helping researchers quickly and accurately discover web technology information with high intelligence value.

[**Method/Process**] Taking into account both external features and textual semantic content features of information, this study utilizes deep learning (pre-trained language model) BERT to construct an information value calculation model based on textual semantic content features. The predictive output of the deep learning model is used for scoring, which is then combined with the original calculation method based on external features to obtain a final comprehensive evaluation score.

[**Result/Conclusion**] Experimental results demonstrate that the information value calculation model based on textual semantic content features can effectively differentiate information into star levels according to their information value scores, compensating for the poor star-level differentiation in the original calculation model based solely on external features. The final comprehensive evaluation results show that the proposed information value calculation model can effectively meet the needs of researchers in practical applications.

Keywords: Web Technology Information; Information Value Calculation; Textual Semantics; BERT

1. Introduction

In the era of big data, the volume of web technology information is growing exponentially, making dynamic monitoring of rich web technology information to support strategic decision-making an increasingly important task for intelligence agencies. However, the massive, multi-source, and complex nature of web technology information presents difficulties and challenges for researchers seeking to timely discover high-value intelligence information. Consequently, how to quickly and accurately identify valuable intelligence content from massive web technology information has become a key focus of information science research.

Through investigation, we found that current calculation methods for web technology information value primarily focus on two aspects: external features of information and user behavior characteristics. The former approach begins with constructing indicator systems, focusing on external attributes of web technology information such as information source, objectivity, and timeliness, using qualitative or quantitative methods for evaluation. The latter approach analyzes the characteristics of user groups from the perspective of information consumers, combining different user preferences to determine the information value of web technology information. Whether based on indicator systems analyzing external features or analyzing user preferences based on behavior characteristics, current research lacks in-depth exploration of the semantic information within the information content itself.

With the development of new technologies such as natural language processing, methods for deep semantic mining of text content have become increasingly mature. Text content is the objective existence form of web technology information, and textual semantic features serve as important references for judging the information value of web technology information. Natural language processing technology has flourished thanks to the emergence of deep learning. The concept of deep learning was proposed by G.E. Hinton et al. in 2006. As an unsupervised feature learning method based on hierarchical feature structures, deep learning simulates the human brain' s neural network for analytical learning, solving many complex pattern recognition problems. Web technology information analysis faces the challenge of massive data, making it highly necessary to explore the application of deep learning technology in this domain.

2. Related Work on Information Value Calculation Methods

Current research on web technology information value calculation methods can be mainly divided into two categories: methods based on external features and methods based on user behavior characteristics.

2.1 Methods Based on External Features

External features of information refer primarily to the external attributes attached to web technology information during production, presentation, and dissemination, such as information source, type, publication time, language, and length. The earliest evaluation indicators for network information were proposed by B. Richmond as the "10C Principles," including content, credibility, critical thinking, copyright, citation, continuity, censorship, connectivity, comparability, and context. Subsequent researchers supplemented these with additional indicators such as information source, text format, comments, timeliness, and originality. With the development of network information technology, evaluation methods based on network link analysis have emerged, most notably the PageRank algorithm proposed by L. Page et al., which calculates the importance of web content by analyzing hyperlink relationships between web pages. The more links associated with a piece of network information, the higher its importance—a principle similar to citation analysis in information science. Similar methods include the HITS algorithm for web page ranking proposed by J.M. Kleinberg.

In recent research, scholars have placed greater emphasis on the scientificity and completeness of indicator system construction. For example, Zhao Yusui et al. applied the Delphi method through expert consultation to establish evaluation indicators for network health information quality, ultimately identifying three first-level indicators (information characteristics, media characteristics, and publication characteristics) and 15 second-level indicators. Deng Shengli et al., from a user perspective, constructed an evaluation framework comprising two first-level indicators (content and design) with seven second-level and seven third-level indicators through user surveys. Liu Jianhua et al. proposed five indicators: information source, information type, relevance to monitoring objects, relevance to science and technology, and relevance to themes, refining them into 31 second-level indicators. These indicators encompass both external features of information resources and some content features, forming a comprehensive evaluation method. This approach was pioneering in extending external features to the thematic level—the textual content dimension of information. We further advance this content-based method by using deep learning to learn contextual features of text content, thereby associating textual content with information value.

2.2 Methods Based on User Behavior Characteristics

The ultimate purpose of dynamic monitoring and analysis of web technology information by intelligence service personnel is to serve users. Whether the provided intelligence services meet target users'information needs determines the effectiveness and quality of information services. Therefore, providing targeted intelligence services by analyzing users' information behavior characteristics is also a key focus of intelligence work. Zhang Yang et al., through a comprehensive review of research on web technology information resource evaluation, proposed

"establishing a user-centered evaluation concept." Many researchers have also explored user behavior characteristics when calculating the information value of web technology information.

As early as 2000, Zhao Jihai included audience as a separate indicator among eight evaluation criteria. Consideration of user behavior characteristics is more evident in resource evaluation and ranking within information retrieval systems. For instance, H. Karodiya et al. classified users of retrieval systems and incorporated user categories to obtain different ranking results for different user groups. Studies by S.L. Price et al., M. Han et al., and L. Tamine-Lechani et al. all explored user interests and preferences to build personalized retrieval services. In recent research, Wang Xiaoli et al. also proposed principles for constructing network information resource evaluation indicators, among which the guiding principle suggests that different users have different needs for network information due to age, cognitive habits, and educational background. Wang Xiwei et al. demonstrated differences in information interaction effectiveness among different network community users, further illustrating that user group characteristics significantly influence the evaluation of network information utilization value.

Users generally pay more attention to the information content itself. Some researchers have built information filtering models based on binary classification, filtering information according to user preferences to provide more valuable intelligence. For example, studies by R. Bing and N. Vatani et al. focused on word features in information content, analyzing word frequency and synonyms to construct user interest models that associate information content with user preferences. We similarly adopt this information filtering approach, linking the content of web technology information texts with user attention by collecting texts considered valuable and non-valuable by users as training sets to build a binary classification model.

3. Research Methods

Many current analytical methods for web technology information value calculation typically construct corresponding indicators based on external features such as information source authority and information type to judge information value. While these external features reflect information value to some extent—for example, information resources from government departments usually have high value and thus high external feature scores—this approach does not deeply explore the semantic meaning of the information text. To address this limitation, we propose an information value calculation model that integrates textual semantic content features, focusing on the textual semantic content level of information. Building upon five external feature dimensions (information source authority, information type, importance of monitoring objects in the information, relevance to science and technology, and thematic relevance), we add the dimension of textual semantic content and integrate all evaluation indicators to obtain the final information value calculation result. The technical route of

the information value calculation model integrating textual semantic content features is shown in Figure 1 [Figure 1: see original paper].

We aim to fully leverage the large-scale unsupervised pre-training of pre-trained language models and the superior text semantic and syntactic feature mining capabilities of Transformer to build an information value calculation model based on textual semantic content features. Annotated corpora are indispensable for supervised learning of machine learning models; however, manual annotation of information value is time-consuming and labor-intensive. Therefore, we first automatically construct the training dataset for information value calculation based on external feature scores to obtain web technology information with and without information value. We then define information value calculation as a binary classification task—predicting whether web technology information has information value or not—and obtain the textual semantic importance score of information resources through the model' s prediction confidence for the valuable category. Finally, we combine the textual semantic importance score with external feature scores to obtain the final information value score.

3.1 Corpus Construction for Information Value Calculation

Leveraging the domain technology intelligence knowledge service cloud platform developed by our project team, we constructed the corpus required for training the information value calculation model. Starting from the needs and workflow of intelligence work, the platform automatically helps intelligence personnel discover the latest and most important technology resources from massive web technology information resources. With the aid of information extraction, automatic classification, automatic summarization, and text mining methods, it automatically calculates important technology objects and important technical terms contained in technology resources—information that is crucial for constructing the corpus for the information value calculation model.

Since domain expert evaluation of information value is time-consuming and labor-intensive, and different experts may have divergent opinions, building a large-scale manually annotated corpus for information value calculation is not feasible. To address this issue, we propose a corpus construction method based on external features of information, fully utilizing five dimensions: information source authority, information type, thematic relevance, authority of monitoring objects, and relevance to science and technology to automatically construct the information value calculation corpus. Specifically, these external features are formulated by intelligence analysts based on empirical knowledge. We believe that external features of information reveal the importance of information resources to a certain extent, and importance thresholds can be set to divide web technology information into valuable and non-valuable categories, initially constructing a large-scale information resource calculation dataset automatically.

The framework of external features for information value calculation is shown in Figure 2 [Figure 2: see original paper]. Based on the above five dimensions,

we can automatically calculate the external feature scores of information value. The energy domain technology information monitoring platform built on the domain technology intelligence knowledge service cloud platform uses external feature scores to measure information importance and feeds back to users. This model has been operating for many years, and users are generally satisfied with the filtered and ranked information. Therefore, we selected reports manually compiled in the platform and reports with external feature importance scores $\geq$ 0.6 as valuable web technology information, and reports without manual compilation as non-valuable web technology information to construct the supervised learning dataset. This indirectly integrates domain ontology, domain subject terms, hot terms, technology subject terms, and important monitoring objects into the model.

Dataset statistics are shown in Table 1 . A total of 22,450 pieces of information were obtained. After random shuffling, they were divided into training and test sets at an 8:2 ratio, with 17,959 pieces in the training set and 4,491 pieces in the test set. The ratio of valuable to non-valuable web technology information in the training set is 9,962:7,997, representing a relatively balanced distribution. The training and test sets contain a total of 12,453 pieces of valuable web technology information and 9,997 pieces of non-valuable information. We assign a label of 1 to valuable web technology information and 0 to non-valuable information to construct the initial corpus for the binary classification model.

3.2 Model Architecture

The proposal of the pre-trained language model BERT in 2018 attracted widespread attention in the natural language processing field. Many researchers have found that using pre-trained language models in natural language processing tasks can significantly improve downstream model performance. The BERT model uses two pre-training tasks on large-scale unlabeled text such as Wikipedia: Masked Language Model (MLM) and Next Sentence Prediction (NSP), learning good general language representations that greatly help improve model performance when transferred to downstream supervised learning tasks. Additionally, Transformer is a superior feature extractor that, through self-attention mechanisms, solves the long-distance dependency problem of long short-term neural networks to some extent and can model text semantics and syntactic features well. We aim to fully leverage the advantages of BERT' s unsupervised pre-training and the Transformer model architecture to build an information value calculation model, while integrating external resource features of information to assist decision-making. We propose an information value calculation model based on textual semantic content, with its architecture shown in Figure 3 [Figure 3: see original paper].

For this model, input information resources are first vectorized, mapping each character in the text to a high-dimensional vector space to obtain character representations. Notably, we add a [CLS] token before each sentence, using this token' s vector representation as the representation of the entire sentence.

The text is then input into a Transformer model stacked with 12 encoder layers to obtain the final vector representation of the [CLS] token, which is fed into a feedforward neural network with SoftMax classification to obtain confidence scores for both non-valuable and valuable categories from the BERT model. We use BERT' s prediction score for the valuable category as the textual semantic content score of the information, which is weighted with external feature scores at a ratio of 0.7:0.3 to obtain the final information value score. Based on this score, thresholds can be set to classify information importance into star levels: when $0.9 \leq$ final information value score ≤ 1, information importance is five-star; when $0.8 \leq$ score < 0.9, four-star; when $0.3 \leq$ score < 0.8, three-star; when $0.1 \leq$ score < 0.3, two-star; and when $0 \leq$ score < 0.1, one-star.

4. Experiments and Results

We selected the energy domain as our experimental field to build and evaluate the information value calculation model. This section introduces the specific experimental procedures and analyzes the results.

4.1 Data Processing

We obtained raw intelligence resources from the energy domain technology information monitoring platform, as shown in Figure 4 [Figure 4: see original paper]. Since intelligence resources are collected and organized semi-automatically, they contain substantial noise data such as special markers and irrelevant webpage titles, which may affect semantic analysis of intelligence text content. To address this noise, we first segment the text into sentences, then develop a series of rules to clean the noise through analysis of the intelligence text. Specific rule examples include:

1. If a sentence contains "加载更多:" or "参考资料:" or "原文出处:" or "推荐阅读:" or "责任编辑:" or "下一篇:" or "上一篇:" or "来源:", delete that sentence;
2. If a sentence starts with "来源:" or "编者按:" or "推荐" or "CAJ 下载" or "PDF 下载." or "HTML 阅读" or "下载频次" or "不支持迅雷" or "免费订阅", delete that sentence;
3. If a sentence contains "发布时间" or "字号" or "来源:", replace "点击收藏" in that sentence with empty;
4. If sentence length < 5, delete that sentence.

After cleaning, we assign label 1 to intelligence with external feature scores ≥ 0.6 and label 0 to intelligence with scores < 0.6, obtaining the dataset for BERT model training. The format is shown in Figure 5 [Figure 5: see original paper]. A total of 17,959 training samples and 4,491 test samples were obtained.

4.2 Experiments and Results

The BERT model achieved an accuracy of 96.77% on the 4,491 test samples. We weighted the BERT model' s information value prediction scores with external

feature scores to obtain final information prediction scores, then classified information into star levels based on these scores. To test the effectiveness of the information value calculation model based on textual semantic content features, we compared it with the model based on external features. Table 2 shows the evaluation thresholds for star levels of both models.

To verify that incorporating textual semantic content features plays an important role in information value calculation, we tested using all 22,450 samples from both training and test sets, counting the distribution of information across star levels. The comparative analysis of statistical results between the two models is shown in Figure 6 [Figure 6: see original paper].

The threshold division for the external feature-based model was set based on the distribution ratio of resource importance detected through testing across multiple domains. As shown in Figure 6, for the energy domain, this threshold concentrates information value prediction scores primarily in one-star and two-star levels, making it impossible to effectively differentiate the information value of web technology information and difficult to select valuable information from massive data. This reflects the domain limitations of the external feature-based information value calculation model. In contrast, the model based on textual semantic content features can effectively differentiate information into star levels, showing better discrimination between valuable and non-valuable web technology information.

Additionally, we conducted a case analysis of one information resource, with the sample shown in Figure 7 [Figure 7: see original paper]. For this resource, the original external feature-based method gave an information value score of 0, while the BERT model gave a prediction score of 0.99997842, and the comprehensive model combining external and textual semantic content features gave a final score of 0.7999. Through analysis of the information text, we found that our method can excavate network resources that were predicted as unimportant by the external feature-based model. While external features such as information source authority can reflect information value to some extent, valuable web technology information may still be hidden in massive network resources. Only by integrating textual semantic features of web technology information and analyzing from the text content itself can information value prediction become more credible.

5. Application Effect Evaluation

We selected the 500 most recent data entries monitored on the energy domain technology information monitoring platform as the test dataset, then applied both the external feature-based information value calculation method and our proposed comprehensive method for scoring. Final star levels were assigned according to their respective star classification standards. The purpose of intelligence is to be utilized, meeting user needs and solving problems. Different domains have different needs and characteristics. Which web technology in-

formation has higher intelligence value should be determined by researchers in that domain. Compared with general researchers, domain experts can more accurately and consistently judge information value. Therefore, evaluating the application effect of domain-specific intelligence value requires evaluation by experts in that domain.

We invited five experts from the Energy Domain Team of the Wuhan Library, Chinese Academy of Sciences, to evaluate the star levels of these 500 data entries, taking the average star rating. These five experts are all users of the energy domain technology information monitoring platform, including three researchers who have long been engaged in advanced energy technology intelligence research and two frontline scientific researchers with doctoral backgrounds. They have urgent needs for accurately identifying valuable energy domain web technology information. The evaluation standard uses four levels to represent the recognition coupling degree between the two calculation methods' results and expert evaluation results: complete recognition (0), relatively recognized (1), relatively unrecognized (2), and completely unrecognized (3). Complete recognition means identical star ratings; relatively recognized means a one-level difference; relatively unrecognized means a two-level or greater difference; completely unrecognized means a three-level or greater difference.

The comparative analysis results are shown in Table 3 . As shown, the external feature-based method achieved 72% relatively recognized or above, while our proposed method achieved 87%, representing a 15% improvement. Therefore, compared with the external feature-based method, our comprehensive information value calculation method can better satisfy researchers in practical applications. However, this evaluation has certain limitations: the expert group size is limited and may not be fully representative; different experts' knowledge levels and subjective needs may cause biased evaluation results; and insufficient test data volume may lead to potential data bias.

6. Conclusion and Future Work

This study comprehensively considers both external features and textual content features of information, utilizes deep learning BERT to construct an information value calculation model based on textual semantic content features, uses deep learning model predictions for scoring, and combines it with the original external feature-based calculation method to obtain final comprehensive evaluation scores. Results show that the textual semantic content-based model can effectively differentiate information into star levels, compensating for the poor differentiation in the original external feature-based model. Compared with methods relying solely on external features, our comprehensive information value calculation method can more effectively identify valuable web technology information and better meet researcher needs in practical applications.

Future research will focus on: 1. **Corpus refinement for information value calculation.** Training set quality determines the practical effectiveness of deep

learning models. We will refine the training corpus, improve corpus construction strategies, iterate based on actual model test results, and form more discriminative datasets of valuable and non-valuable web technology information. 2. **Expanding application domains.** We will attempt to construct domain-specific information value calculation models according to the linguistic characteristics and user needs of different disciplines.

References

[1] Zhang Zhixiong, Zhang Xiaolin, Liu Jianhua, et al. Implementation of ideas and technical methods for structured monitoring of web technology information[J]. Journal of Library Science in China, 2014, 40(4): 4-15. [2] Zou Yimin. Research on information value judgment method based on object calculation[D]. Beijing: University of Chinese Academy of Sciences, 2013. [3] Zhang Yang, Zhang Lei. Review of research on network information resource evaluation[J]. Journal of Library Science in China, 2010, 36(5): 75-89. [4] Zou Yimin, Zhang Zhixiong. Review of web technology information value evaluation methods[J]. Journal of Intelligence, 2014, 33(5): 25-30, 59. [5] Hinton G E, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets[J]. Neural Computation, 2006, 18(7): 1527-1554. [6] Richmond B. 10Cs for evaluating Internet resources[C]//Proceedings of the 17th international ess symposium. Washington: ERIC, 1994: 287-312. [7] Stoker D, Cooke A. Evaluation of networked information sources[J]. Public access computer systems review, 1997, 8(3): 1-14. [8] Smith A G. Criteria for evaluating information resources[J]. Teacher librarian, 1998, 25(5): 20. [9] Zhao Jihai. Internet information evaluation: an important responsibility for librarians in the new century[J]. Journal of Academic Libraries, 2000(5): 35-38. [10] Su Guangli. Research on evaluation of Internet information resources[J]. Information and Documentation Services, 2001(6): 26-28. [11] Page L, Brin S, Motwani R, et al. The pagerank citation ranking: bringing order to the Web[R]. Stanford: Stanford InfoLab, 1999. [12] Kleinberg J M. Authoritative sources in a hyperlinked environment[J]. Journal of the ACM, 1999, 46(5): 604-632. [13] Zhao Yusui, Xu Yan, Wu Qingqing, et al. Application of Delphi method to construct evaluation index system for network health information quality[J]. Preventive Medicine, 2018, 30(2): 121-124. [14] Deng Shengli, Zhao Haiping. Research on construction of evaluation standard framework of network health information quality from user perspective[J]. Library and Information Service, 2017, 61(21): 30-39. [15] Liu Jianhua, Zhang Zhixiong. Indicator system and calculation method of information importance[R]. Beijing: National Science Library, Chinese Academy of Sciences, 2011. [16] Karodiya H, Singh A P D K. User specific search ranking technique[J]. International research journal of computer science engineering and applications, 2013, 2(1): 212-215. [17] Prince S L, Nielsen M L, Delcambre L, et al. Using semantic components to search for domain-specific documents: an evaluation from the system perspective and the user perspective[J]. Information systems, 2009, 34(8): 724-752. [18] Han M, Qiu X H. Personalized search engine model[C]//Advanced materials research. Switzerland: Trans Tech Publications Ltd, 2011: 1216-1221. [19]

Tamine-Lechani I L, Boughanem M, Zemirli. Personalized document ranking: exploiting evidence from multiple user interests for profiling and retrieval[J]. Journal of digital information management, 2008, 6(5): 354-366. [20] Wang Xiaoli, Yan Shi, Liu Zhanbo, et al. Analysis of construction of network information resource evaluation index system[J]. Software, 2020, 41(5): 53-56. [21] Wang Xiwei, Zhang Changliang, Han Xuewen, et al. Research on evaluation of information interaction effect of network community from the perspective of information ecology[J]. Information Studies: Theory & Application, 2018, 41(11): 83-88, 62. [22] Bing R. Information filtering algorithm based on feature vector[C]//Proceedings of the 2011 international conference on intelligence science and information engineering. New York: IEEE, 2011: 468-471. [23] Vatani N, Shirie M. A personalized information filtering method[J]. International journal of computer science and security, 2012, 6(1): 1-8. [24] Devlin J, Chang M W, Lee K, et al. Bert: pre-training of deep bidirectional transformers for language understanding[EB/OL]. [2020-12-25]. https://arxiv.org/abs/1810.04805. [25] Beltagy I, Lo K, Cohan A. Scibert: a pretrained language model for scientific text[EB/OL]. [2020-12-30]. https://arxiv.org/abs/1903.10676. [26] Lee J, Yoon W, Kim S, et al. Biobert: a pre-trained biomedical language representation model for biomedical text mining[J]. Bioinformatics, 2020, 36(4): 1234-1240.

Author Contributions

Zhang Min: Conceptual design, manuscript writing and finalization;
Liu Huan: Experimental verification and manuscript proofreading;
Ding Liangping: Data organization and manuscript proofreading;
Fan Qing: Method application and evaluation.

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