Unsupervised Tagging of Chinese Articles

Shailendra Singh Kathait
Co-founder & Head
Artificial Intelligence & Machine Learning Lab
Valiance Solutions, Noida, Uttar Pradesh, 201301

Shubhrita Tiwari
Data Scientist Artificial Intelligence & Machine Learning Lab Valiance Solutions, Noida, Uttar Pradesh, 201301

ABSTRACT
Large amount of insights can be drawn from the articles that are published online. Instead of manually reading all the articles and assigning relevant tags to them satisfying the content, it will be highly efficient if there exists an automated process for performing the task. In this paper, an unsupervised approach for the automated tagging of articles in Chinese language has been implemented. The input is an article and output is the tags to that article. The major challenge is the segmentation of the Chinese characters, which do not make use of separators unlike the English characters. To overcome this, different approaches are combined together in order to get accurate results. Efficient tagging of articles is required, which can be used for many applications in the analysis, one of which is in Recommendation Engine. The tagging process should consider all the aspects of the article and assign the most relevant tags accordingly. The proposed algorithm was implemented for a Chinese Publication House and relevant tags were assigned to its articles of different categories. At the end of the project, the results were manually checked for, in a corpus of 10000 Chinese articles, which reflected the attainment of overall accuracy of around 85%, greater than that obtained through different traditional methods.

Keywords
Text, articles, automated tagging, tags, unsupervised approach, Recommendation Engine, Chinese, segmentation, corpus.

1. INTRODUCTION
The tags assigned to the articles give a rough picture of the content of the article, without reading it completely. In this world of data, where there are lots of articles published online, automated tagging has got much significance. To retrieve information or to give results for a particular query of user, automated tagging is required. Manually tagging or summarizing such articles will be highly inconvenient, and this calls for automation so as to reduce the time and effort. Article tagging plays an important role in search engine queries, text classification and summarization, Recommendation engine etc.

The tagging process starts with the key-phrase extraction that can be further filtered for relevant tags. The aim of key-phrase extraction model is to generate words and phrases that would together summarize the entire text.

The algorithms for key-phrase extraction can be broadly classified into two types [1][2] –

- Supervised key-phrase extraction
- Unsupervised key-phrase extraction

Supervised method employs choosing the best keywords from a prepared set of keywords, which is likely to contain topics from all genres and fields of interest. This method requires labeled articles with their tags or keywords. A model is developed to learn the ways in which the tags and keywords can be associated with an article and how they can be generated from an article. While this can produce interpretable rules as to what characterizes a key-phrase, but the greatest challenge in this method is the availability of training data, tags and keywords for large number of texts whose topics encompass all genres of interest like scientific journals, news, business, education, entertainment, sports etc. Also, this method would not be able to incorporate the actual context of the text or in other words it will lack specificity.

The unsupervised method eliminates the need for labeled training data. Unlike supervised methods, the unsupervised method uses the structure of the article itself and generates keywords and phrases from the article using its properties. This approach holds undeniable importance as this can be applied across all languages and domains.

In order to assign tags to Chinese articles, a novel unsupervised approach is used that gives the most relevant tags to the articles, that can be later used to generate recommendations.

The major challenge in the entire process is the segmentation of the Chinese characters. Since an unsupervised approach is being used, the segmentation plays an important role in the assignment of tags. The tagging of Chinese articles is much more difficult as compared to its English counterpart, since Chinese texts require more complex algorithm to perform word-segmentation.

The major challenges for analyzing Chinese texts are:

1. Separation of English and other irrelevant content from Chinese text.
2. Segmentation of Chinese characters.
3. Performing the matching operation of the segmented words with those in predefined dictionary.
4. Preparation of a predefined dictionary that contains all the stop words that is to be removed from the text.
5. Preparation of predefined dictionary that contains the common tags to be assigned to the articles of specific fields.

The method used in this paper is unsupervised one, in which no supervision and outside interference is required once the algorithm is built. After the segmentation of Chinese characters, the relevant words are selected as tags for the articles.

2. RELATED WORKS
Different approaches have been used to generate keywords from the texts. In [3], HMM based approach is used to generate tags for the articles, which does not cover articles of all categories. In supervised methods used in [4], the same problem remains of covering all types of articles of different fields in the training-set. In approach adopted in [5], the desired accuracy for all types of articles was not achieved. In ontology based approach in [6], a limited number of

predefined tags are assigned to the articles, which does not give the most accurate results. All the aforementioned methods have been used for articles in English, which is comparatively easier since English words are separated by space.

We used a novel unsupervised approach to assign the most accurate tags to the Chinese articles so as to cover all types of articles of different fields.

3. PROPOSED METHODOLOGY

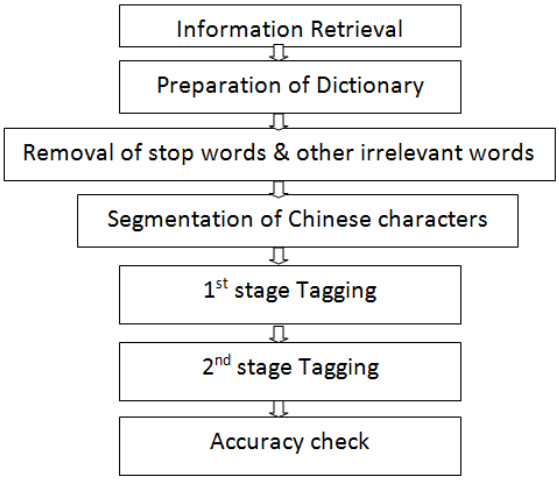


Figure-1 Proposed Methodology

3.1 Information Retrieval:

The first step in the entire process is fetching of data and collecting into database. The input database consists of different articles in Chinese with their corresponding Ids and content.

3.2 Preparation of Dictionary:

3 different dictionaries are required:

- One dictionary of Proper Nouns, which consist of all the common, popular and trending Proper Nouns that should be assigned as tags, if present in the article. This is prepared manually by using articles provided by publishers.
- Second one is the data of stop words that will not contribute to tagging. This is also prepared manually.
- The third one is the master dictionary that keeps the result obtained after performing TF-IDF operations in addition to some tags assigned manually.

3.3 Removal of irrelevant words:

The Articles consist of different html tags in English etc and other irrelevant stuff that require cleaning, since they do not contribute to tagging.

3.4 Removal of stop words:

Stop words are the irrelevant words that are filtered out, since they do not contribute to tagging (Example: Articles a, an, the etc.). A predefined dictionary of stop words, prepared in step-2, is used here.

3.5 Segmentation of Chinese Characters:

The different approaches to the segmentation process are:

3.5.1 Maximum probability method:

Using Search tree method, directed acyclic graphs are constructed. Using memory search, the maximum probability path is calculated. Word graph scanning is done and search algorithm is run in the predefined dictionary.

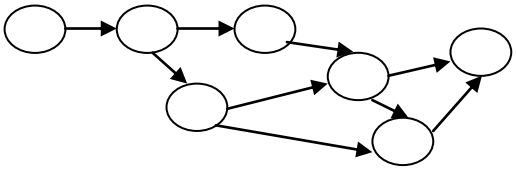
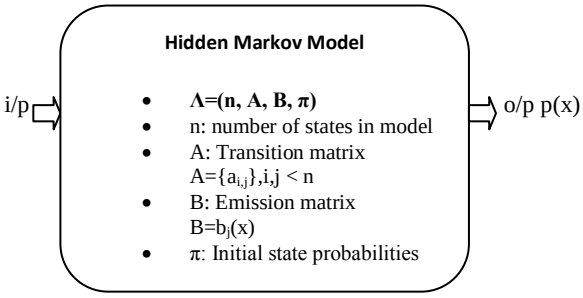


Figure-2 Directed Acyclic Graphs

3.5.2 Hidden Markov method

This is used for words not present in the predefined dictionary. It uses Hidden Markov model to determine status set and observed set of words. The default Hidden Markov model is based on People's Daily language library. In this Viterbi Algorithm is used for segmentation.



3.5.3 Mix Segment method

This makes use of both Maximum probability segmentation method and Hidden Markov method to construct segmentation output.

Input: 中国文章的标签很难

Output: 中,国, 文,章, 的, 标,签, 很,难

After the relevant segmented Chinese characters are extracted, the tagging process is started.

3.6 1st Stage Tagging:

The preprocessed and filtered words (after removal of stop words and other irrelevant words) are now searched for in the Proper Nouns dictionary and if the words are present in the dictionary, then the corresponding tags of particular category, are assigned to the articles.

3.7 2nd Stage Tagging:

The segmented words are searched for in the predefined custom dictionary. If not present, they are appended into the dictionary. The TF-IDF is implemented and certain weight is assigned to each word. If the weights lie within a certain threshold value, they are chosen as tags and assigned to the articles containing them.

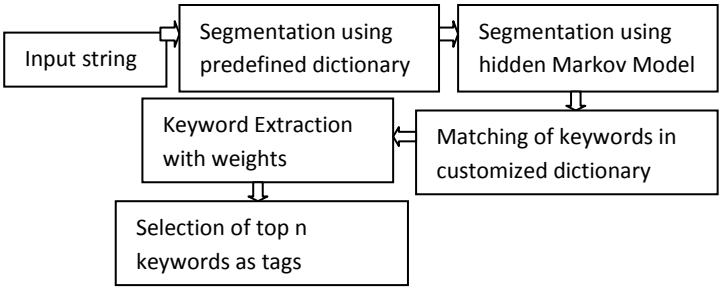


Figure-3 Tagging Flow Diagram

The following concepts are used in tagging:

Term Frequency (TF): Term Frequency (TF) of term t in document d is defined as the number of times that t occurs in d. That is, the separated n-grams are searched for in the single article and the number of times, it is found is termed as term frequency.

$$\text{TF (t)} = \frac{Number\ of\ times\ term\ t\ appears\ in\ a\ document}{Total\ number\ of\ terms\ in\ the\ document} \quad (1)$$

Inverse Document Frequency (IDF): Estimate the rarity of a term in the whole document collection. (If a term occurs in all the documents of the collection, its IDF is zero.). In our case, the n-gram terms are searched for in the corpus with large number of Chinese articles, and IDF is given by the inverse of this number. If IDF comes out to be lying within a certain lower threshold range, then those terms are selected and passed on to the next filter [7].

The mathematical expression used is:

$$\text{IDF} = \log |D| \, |\{j : t \in d\}| \quad (2)$$

Where |D| : cardinality of D, or the total number of documents in the corpus

$|\{j : t \in d\}|$: number of documents where the term t appears (viz. the document frequency).

If the term is not in the corpus, this will lead to a division-by-zero. It is therefore commonly used with 1+.

Combining the definitions of term frequency and inverse document frequency, produces composite weight for the segmented Chinese characters .The TF-IDF weighting scheme assigns weight to the terms as:

$$\text{TF-IDF= TF.IDF} \quad (3)$$

The concept of TF-IDF along with n-gram modeling is used to prepare Master dictionary that consist of the words obtained as a result of TF-IDF.

The TF-IDF calculated above assigns certain weights and if it falls under a certain range of threshold values, then it is searched for in the Master dictionary (that contains different tags of numerous categories). If it is present in the Master Dictionary, then the corresponding tags are returned and if it is not present in the dictionary, then it is returned as tag to the article and appended in the Master dictionary as well.

4. IMPLEMENTATION

The following steps are involved in the tagging process:

- The database of 10000 articles in Chinese language is fetched.
- Using TF-IDF concept, n-gram modeling and some manual addition, Master Dictionary is prepared with the relevant key-words.
- Dictionary containing stop words and Proper Nouns is also prepared manually.
- Segmentation of the Chinese characters is done by using Maximum probability method and Hidden Markov method along with use of Master Dictionary.
- After the word-segmentation is completed, the keywords are searched for in the Master dictionary, assigning certain weights to the keywords (these weights are assigned on the basis of TF-IDF), from which top n keywords are selected as tags, where n can be any value between 5-10 depending upon the use case.

Example:

Article Content-

由于美元淹入当地股票和债券，周三卢布的全球货币走势大幅上涨至20个月以来的最高点，促使印度央行干预以限制收益，从而削弱了出口商的盈利能力。

经纪商表示，卢比的销售看起来是由中央银行引发的，从日前的高点回升，但印度货币仍然以64.11的价格上涨了0.25%，触及了美元的心理重要标志。它触及了高点6。

Relevant Tags generated-

卢比，全球，货币，印度，出口商，销售

Irrelevant Tags-

市场，国家，娱乐

5. OBSERVATIONS

- The proposed algorithm was first tested for accuracy on single article of different categories.
- The tags of each article are manually checked for accuracy.
- Accuracy for each article is defined as:

$$\text{Accuracy} = \frac{Number\ of\ relevant\ tags}{Total\ number\ of\ tags} \quad (4)$$

The observed results for single article of each category are tabulated as:

Table-1 OBSERVATION TABLE

Type of Article	Number of relevant tags	Number of irrelevant tags	Accuracy
Banking	11	2	0.85
Insurance	8	2	0.80
Financial	7	3	0.70
Education	11	2	0.85
Entertainment	6	4	0.60

- The algorithm is further implemented for 10000 articles of different categories, which were chosen randomly from a corpus of 100000 articles.
- Overall accuracy is taken as average of the accuracies obtained for each article, given as:

$$\text{Overall Accuracy} = \frac{\sum Accuracy\ for\ each\ article}{10000} \quad (5)$$

- The overall accuracy was obtained close to 85%, when calculated for articles of different categories using the aforementioned formulae.

6. CONCLUSION

The algorithm implemented gave accurate results, as can be seen from the observation table (Table-1) but still there lies a scope of improvement, especially in some categories of articles like Entertainment for which accuracy is minimum.

For these types of articles, supervised learning approach can be used in-order to get accurate results. A novel features based approach based on citation network information used in conjunction with traditional features for key-phrase extraction and then using those keywords as tags, can be implemented that gives remarkable improvements in performance over strong baselines [8].

7. FUTURE WORK

The supervised methods can be tested for more accurate results. Addition of more articles in the data from different categories as well as more words in the predefined dictionary can further increase the accuracy.

8. REFERENCES

[1] Aditi Sharan, Siddiqi, Sifatullah, "Keyword and key-phrase extraction techniques: A literature review", International Journal of Computer Applications 109, no. 2 (2015).

[2] Ana, Beliga, Sanda, Slobodan, "An overview of graph-based keyword extraction methods and approaches", Journal of Information and Organizational Sciences 39, no. 1 (2015).

[3] Chao-Huang Chang, Cheng-Der Chen, "HMM-based Part-of-Speech Tagging for Chinese Corpora", Hsinchu, Taiwan, R.O.C.

[4] Ben Taskar, Joao V. Graca, Shen Li, "Wiki-ly Supervised Part-of-Speech Tagging", University of Pennsylvania.

[5] "Automatic free-text-tagging of online news archives".

[6] Gridaphat Sriharee, "An ontology based approach to auto-tagging articles", University of Technology, North Bangkok, Bangkok, Thailand.

[7] Tf-idf weighting - Stanford NLP Group https://nlp.stanford.edu/IR-book/html/htmledition/tf-idf-weighting-1.html.

[8] Andreea Godea, Cornelia Caragea, Florin Bulgarov, Sujatha Das Gollapalli, "Citation-Enhanced Keyphrase Extraction from Research Papers: A Supervised Approach", Singapore.