Extraction of Historical Transitions in Legal and Scientific Laws from Wikipedia

Liangliang Fan[1], Masaki Murata[1], Masato Tokuhisa[1] and Qing Ma[2]
[1]*Department of Information and Electronics, Tottori University*
Tottori, 680-8550 Japan
[2]*Department of Applied Mathematics and Informatics, Ryukoku University*
Otsu, 520-2194 Japan

[1]{*k112001;murata;tokuhisa*}*@ike.tottori-u.ac.jp*
[2]*qma@math.ryukoku.ac.jp*

In recent years, the free online encyclopedia Wikipedia has been widely used worldwide. It would be convenient if the historical transitions in laws (including legal, mathematical and physical laws), which expresses a change in the laws, can be automatically extracted from Wikipedia. In this paper, we propose a new method of extracting the historical transitions in laws with heuristic-based and machine-learning-based methods. We use the machine-learning-based method for improving the performance of the heuristic-based extraction of the historical transitions in laws. Using only the heuristic-based method in our experiment, we achieved an F-measure of 0.46. On the other hand, using the machine-learning-based method in addition to the information used for the heuristic rule, we achieved an F-measure of 0.68. In the experiments that involved extracting the law pairs that have a transitional relationship without extracting the law years, we achieved an F-measure of 0.87 using the machine learning method.

Keywords: transition; law; heuristic; feature; machine learning.

1. Introduction

In recent years, the free online encyclopedia Wikipedia has been widely used worldwide[1]. It would be convenient to be able to automatically extract the historical transitions, which show their evolution over time, of laws (including legal, mathematical and physical laws) from Wikipedia, and we aim to achieve this goal in our study[a].

In this paper, we define the historical transitions in laws to include law pairs that have a transitional relationship and the year (hereinafter referred to as the law year) in which the law was discovered. Some examples of the historical transitions in laws are shown in Table 1. For example, "Game theory" was derived from "Decision theory".

[a]The significance of the extraction of the historical transitions in laws includes the following: Historical transition in laws is basic information of a law, and it is convenient if it can be collected and arranged automatically. It becomes easier to understand the relation between laws. Moreover, it is useful to arrange the history of scientific development like Hori et al.'s study[2].

Table 1. Examples of historical transitions in laws

Law A	Law B
Decision theory (1670)	Game theory (1928)
Decision theory (1670)	Prospect theory (1979)

In this study, the method used to extract the historical transitions in laws is as follows: For example, based on the heuristic rule, when a law B is described in the pages of a given law A, law B is more likely to have a transitional relationship with law A. Therefore, we extract a law pair comprising law A and law B as historical transitions information, together with the discovery year of each law, which is determined by the years that have been described in the pages of law A (or B). The heuristics-based method for extraction of law years is to extract the year at the top of a page of law A (or B), which is likely to be the discovery year of the law A (or B). To improve the performance of the heuristic-based method, we also used supervised machine learning.

We performed this study in the Japanese language.

The main points proposed in this paper are as follows:

- We proposed a heuristic-based method that extracts the year at the top of the law page as the law year and extracts a basic law and its related law as law pairs that have a transitional relationship. Using this simple method, we were able to achieve a historical transitions with an F-measure of 0.46.
- In addition to the heuristic rule mentioned above, we proposed a machine learning method. Using this method, the performance was significantly improved, and the historical transitions in laws was achieved with an F-measure of 0.68.
- For the extraction of law pairs that have transitional relationships (in the cases where law years do not need to be extracted), a high F-measure of 0.87 was obtained using the machine learning method.
- The proposed method can be applied to problems that have structures similar to this subject. For example, it can be applied to acquire historical transitions using the discovery year in a certain group of pages that are related to topics other than laws present on Wikipedia.

2. Related work

In a study of information extraction, Hori et al.[2] automatically extracted historical transitions in researcher and research fields using co-occurrence information. Sumida et al.[3] extracted a large quantity of hyponymy information from the section heading and bullet points contained in a Wikipedia article and developed a method to precisely acquire hyponymy information using machine learning. Nakayama et al.[4]

proposed an efficient link mining method pfibf (Path Frequency Inversed Backward link Frequency) and the extension method "forward / backward link weighting (FB weighting)" to construct a large scale association thesaurus. Matsumoto et al.[5] proposed a method to estimate emotion from sentences that include Wakamono Kotoba by using statistical learning methods such as Naive Bayes method and Accumulation method. Quan et al.[6] made an analysis on sentence emotion based on emotion words using Ren-CECps (a Chinese emotion corpus). By comparing some classification methods (including C4.5 decision tree, SVM, NaiveBayes, ZEROR, and DecisionTable), they proposed a supervised machine learning method (Polynomial kernel method) to recognize the eight basic emotions (Expect, Joy, Love, Surprise, Anxiety, Sorrow, Angry and Hate). Ren[7] discussed the definition, intension, and extension of language engineering, affective computing, and advanced intelligence, as well as the relationship among the three fields. Many other studies have performed information extraction from Wikipedia[8,9,10,11]. However, the extraction of historical transitions in laws has not been addressed.

3. Methods

We proposed a heuristic-based method and a machine-learning-based method as methods to extract the historical transitions in laws. For the machine-learning-based method, a support vector machine (SVM) that had excellent performance was used (the secondary polynomial kernel is used for a kernel function[b]).

In our proposed methods, the extraction of historical transitions in laws is performed by extracting law years and law pairs. Details of the extraction of law years and law pairs are described below. A flow diagram illustrating the extraction of historical transitions in laws is shown in Figure 1.

3.1. *Extraction of Law Years*

We extracted the discovery year of the law from pages in Wikipedia whose title is the name of the law. The discovery year of the law is often described in the law page of Wikipedia. Therefore, we extracted the law years from the law pages.

Three methods for extracting law years from the law pages are shown below. Method A1 is a method based on a heuristic rule, and Methods A2 and A3 are based on machine learning.

Method A1 A1 is a method that outputs the first year of a law page as the discovery year of the law. This is because the year appearing at the beginning of the law page often corresponds to the discovery year of the law. The first year of the law page that is extracted as the law year is the output of this method.

[b]We confirmed that the secondary polynomial kernel achieved better performance than the linear kernel in the experiments.

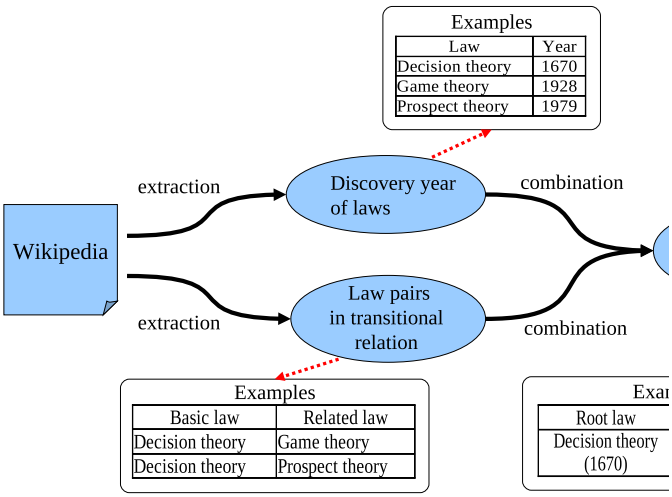


Fig. 1. Flow of the extraction of historical transitions in laws

Method A2 A2 is a method that extracts the first year of a law page, and using machine learning it determines whether the extracted year is the discovery year of the law. Unlike Method A1, in Method A2, the first year of the law page is not directly outputted as the law year. Instead, it is determined to be correct or incorrect using machine learning, and only the correct one is outputted.

Method A3 A3 is a method that extracts all the years described in the law page and gives a score[c] for each year by machine learning. The year with the highest score is the output of this method. In the case where the year with

[c]We used a distance from the separating plane in SVM as the score.

Table 2. Features used in Method A2

Features	Content
f1	Character strings around the year
f2	Position of the year in the page

Table 3. Features used in Method A3

Features	Content
f1	Character strings around the year
f2	Position of the year in the page
f3	Occurrence order of the year in the page

Table 4. Examples of A.D. conversion

	A.D.
34 *Showa* (Japanese calendar)	1959
1000 BC	-1000

the highest score has a negative score (meaning the incorrect year), it is not outputted.

For the extraction of law years, the features used in machine learning (Methods A2 and A3) are shown in Tables 2 and 3.

A.D. conversion is performed when the extracted year is not in the A.D. format. The A.D. conversion confirms whether a law year is in the A.D. format, and when it is not, a program automatically changes the law year into the A.D. format. An example of A.D. conversion is shown in Table 4.

3.2. *Extraction of Law Pairs*

In this study, the law used as the title of a law page is called a basic law, and other laws that exist in the law page are called related laws. Law pairs that have a transitional relationship are called law pairs of a root law and its derivation law, and the law in the pair with the earlier law year then becomes a root law. Using the example of Table I, because the law pair "Decision theory (1670) and Game theory (1928)" is information that has a historical transition, it becomes a pair of a root law and a derivation law. Because the discovery year (1670) of the law "Decision theory" precedes the discovery year (1928) of the law "Game theory", the law "Decision theory" becomes the root law.

Many of the pairs comprising a basic law and a related law that are extracted

Table 5. Features used in Method B2

Features	Content
f1	Name similarity of the law pair
f2	Whether the law pair is a bidirectional law pair

Table 6. Example of a bidirectional law pair

Organic law	Related law
Law C: Game theory	Law D: Decision theory
Law D: Decision theory	Law C: Game theory

from the law pages have a transitional relationship. Therefore, in the extraction of law pairs, a pair that has a transitional relationship, i.e., the pair of a root law and a derivation law, is extracted from the pair of a basic law and its related law.

Two methods of extracting the pair of a root law and its derivation law from the law pages are shown below. Method B1 is a method based on a heuristic rule, and Method B2 is a method based on machine learning.

Method B1 B1 is a method that considers all the pairs of a basic law and its related laws extracted from the law pages to have a transitional relationship.

Method B2 B2 is a method that uses machine learning to determine whether a transitional relationship exists between the pair of a basic law and its related law extracted from the law page. When machine learning determines that the extracted law pair does not have a transitional relationship, it is not outputted. However, when it is determined to have a transitional relationship, the law pair is outputted.

For the extraction of law pairs, the features used in machine learning (Method B2) are shown in Table 5.

The bidirectional law pair that was shown in Table 5 is defined below. For the pair of a certain law C and law D, when law D is described in the page of law C, and law C is conversely described in the page of law D, this pair is called a bidirectional law pair. An example of a bidirectional law pair is shown in Table 6.

3.3. *Extraction of Historical Transitions in Laws*

We extracted the historical transitions in laws by combining a law pair that has a transitional relationship using the method in Section 3.2, with the discovery year of a law that was extracted using the method in Section 3.1. To extract the historical transitions in laws, we use the following six methods by combining three methods for the extraction of law years (Section 3.1) and two methods for the extraction of law pairs (Section 3.2).

Table 7. Results of the extraction of law years

Method	Recall	Precision	F-measure
Method A1	0.92	0.61	0.74
Method A2	0.76	0.85	0.80
Method A3	0.68	0.83	0.75

Method C1 Combination of Method A1 and Method B1
Method C2 Combination of Method A2 and Method B1
Method C3 Combination of Method A3 and Method B1
Method C4 Combination of Method A1 and Method B2
Method C5 Combination of Method A2 and Method B2
Method C6 Combination of Method A3 and Method B2

4. Experiments

4.1. *Experimental Data*

In the experiments, we used pages that were downloaded from the Japanese Wikipedia site on May 26, 2010. We downloaded 1,634 pages and extracted a total of 2,074 law pairs (consisting of a basic law and its related law), 1,621 of which were unique to these pages. The law pages were extracted using manual checks after being taken out by a pattern.

From the 1,621 different law pairs, we randomly extracted 100 law pairs as Data set A (different laws: 133), and without overlapping with Data set A, we randomly extracted another 100 law pairs as Data set B (different laws: 137).

4.2. *Calculation of performance*

In this experiment, we use the recall, precision, and F-measure as parameters that indicate the performance of the extraction.

4.3. *Extraction of Law Years*

In the experiments of using machine learning, we used the 133 laws of Data set A as training data and 137 laws of Data set B as test data. The results of the experiment are shown in Table 7.

4.4. *Extraction of Law Pairs*

In the experiments of using machine learning, we used the 100 law pairs of Data set A as training data and 100 law pairs of Data set B as test data. The results of the experiment are shown in Table 8.

Table 8. Results of the extraction of law pairs

Method	Recall	Precision	F-measure
Method B1	1.00	0.60	0.75
Method B2	0.87	0.88	0.87

4.5. *Extraction of Historical Transitions in Laws*

We set three criteria to evaluate the extraction accuracy of the historical transitions in laws.

Criterion 1 The output of a law pair is judged to be correct when the law pair has an actual transitional relationship and the law years of the pair are correct.

Criterion 2 The output of a law pair is judged to be correct when the law pair has an actual transitional relationship, and the correct order of the two laws can be presumed by the two law years of the law pair, even if the law years of the pair are partially incorrect.

Criterion 3 The output of a law pair is judged to be correct when the law pair has an actual transitional relationship (i.e., the law years do not need to be properly extracted).

In the experiments of using machine learning, we used the 133 laws and 100 law pairs of Data set A as training data and 100 law pairs of Data set B as test data. The results that applied the three criteria to each extraction method for the historical transitions in laws are shown in Table 9. Extracted examples of the historical transitions in laws are shown in Table 10.

5. Discussion

From Tables 7 and 8, we can see that the methods for the extraction of law years described in Section 3.1 and the extraction of law pairs described in Section 3.2 can generally result in high F-measures ranging from 0.7 to 0.8. In particular, we obtained a high F-measure of 0.87 when using the machine learning method for the extraction of law pairs.

Then, we discuss the performance of the extraction of the historical transitions in laws from Table 9.

First, an F-measure of 0.46 was achieved by the method (Method C1) based on the heuristic rule. We can easily realize this performance using the heuristic-based method, which extracts the year at the top of the law page as the law year and extracts the pairs of a basic law and its related law as law pairs that have a transitional relationship.

Second, the accuracy of the extraction of the historical transitions in laws may be improved using machine learning with an F-measure of 0.68. We found that it is possible to improve the performance of the extraction of the historical transitions

Table 9. Results of the extraction of historical transitions in laws

Method	Criteria	Recall	Precision	F-measure
C1	1	0.97 (30/31)	0.30 (30/100)	0.46
	2	0.97 (30/31)	0.30 (30/100)	0.46
	3	1.00 (60/60)	0.60 (60/100)	0.75
C2	1	0.71 (22/31)	0.65 (22/ 34)	0.68
	2	0.71 (22/31)	0.65 (22/ 34)	0.68
	3	1.00 (60/60)	0.60 (60/100)	0.75
C3	1	0.48 (15/31)	0.60 (15/ 25)	0.54
	2	0.52 (16/31)	0.64 (16/ 25)	0.57
	3	1.00 (60/60)	0.60 (60/100)	0.75
C4	1	0.71 (24/34)	0.41 (24/ 59)	0.52
	2	0.71 (24/34)	0.41 (24/ 59)	0.52
	3	0.87 (52/60)	0.88 (52/ 59)	0.87
C5	1	0.50 (17/34)	0.71 (17/ 24)	0.59
	2	0.50 (17/34)	0.71 (17/ 24)	0.59
	3	0.87 (52/60)	0.88 (52/ 59)	0.87
C6	1	0.35 (12/34)	0.71 (12/ 17)	0.47
	2	0.35 (12/34)	0.71 (12/ 17)	0.47
	3	0.87 (52/60)	0.88 (52/ 59)	0.87

Table 10. Examples of historical transitions in laws

Root law	Derivation law
Decision theory (1670)	Game theory (1928)
Decision theory (1670)	Prospect theory (1979)
Hofmann rearrangement (1871)	Curtius rearrangement (1890)
Handenshuju law (646)	Sanseiisshin law (723)
Lex Duodecim Tabularum (451 BC)	Lex Hortensia (287 BC)
Lex Duodecim Tabularum (451 BC)	Lex Licinia Sextia (367 BC)

in laws using a machine learning method (Method C2) in addition to the heuristic rule mentioned above.

Third, in the extraction of law pairs that have the transitional relationship of Criterion 3 (the law years do not need to be extracted), a high F-measure of 0.87 was obtained using a machine learning method (Methods C4, C5, and C6).

6. Conclusion

In this study, we proposed a method for extracting the historical transitions in laws from Wikipedia. The extraction of historical transitions in laws was performed

using heuristic-based and machine learning methods. The heuristic-based method extracts the year that is present at the top of the law page as the law year and extracts the pair of a basic law and its related law as a law pair that has a transitional relationship. From the experiments, using just the heuristic-based method gives an F-measure of 0.46. Using the machine learning method in addition to the information used for the heuristic rule, we achieved an F-measure of 0.68. In the experiments for extracting the law pair that has a transitional relationship without extracting law years, we achieved an F-measure of 0.87 using the machine learning method.

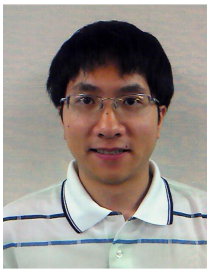
Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 23500178.

References

1. Wikipedia: http://ja.wikipedia.org/wiki/
2. S. Hori, M. Murata, M. Tokuhisa, Q. Ma. Automatic Extraction of Historical Transitions in Researchers and Research Topics, In *Proceedings of the 7th. IEEE Conference on Natural Language Processing and Knowledge Engineering*, pp.296-299, 2011.
3. A. Sumida, K. Torisawa. Hacking Wikipedia for Hyponymy Relation Acquisition, In *Proceedings of the Third International Joint Conference on Natural Language Processing* (IJCNLP2008), pp.883-888, 2008.
4. K. Nakayama, T. Hara, S. Nishio. Wikipedia mining for an association web thesaurus construction, *WISE'07 Proceedings of the 8th International Conference on Web Information Systems Engineering*, pp.322-334, 2007.
5. K. Matsumoto, Y. Konishi, H. Sayama and F. Ren. Analysis of Wakamono Kotoba Emotion Corpus and Its Application in Emotion Estimation, *International Journal of Advanced Intelligence*, pp.1-24, March, 2011.
6. C. Quan, F. Ren. Sentence Emotion Analysis and Recognition Based on Emotion Words Using Ren-CECps, *International Journal of Advanced Intelligence*, pp.105-117, July, 2010.
7. F. Ren. From Cloud Computing to Language Engineering, Affective Computing and Advanced Intelligence, *International Journal of Advanced Intelligence*, pp.1-14, July, 2010.
8. F. Wu, R. Hoffmann, D. Weld. Information Extraction from Wikipedia: Moving Down the Long Tail, *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2008.
9. Y. Yan, N. Okazaki, Y. Matsuo, Z. Yang, M. Ishizuka. Unsupervised Relation Extraction by Mining Wikipedia Texts Using Information from the Web, In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pp.1021-1029, 2009.
10. D. Nguyen, Y. Matsuo, M. Ishizuka. Relation Extraction from Wikipedia Using Subtree Mining, *AAAI'07 Proceedings of the 22nd National Conference on Artificial intelligence*, pp.1414-1420, 2007.
11. I. Dagan, L. Barak, E. Shnarch. Extracting Lexical Reference Rules from Wikipedia, *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP*, pp.450-458, 2009.

Liangliang Fan



He is an M.S. student in the Department of Information and Electronics, Tottori University, Japan. He now focuses on information extraction.

Masaki Murata

He received his Bachelor's, Master's, and Doctorate degrees in engineering from Kyoto University in 1993, 1995, and 1997, respectively. He worked in the Communications Research Laboratory (Now: National Institute of Information and Communications Technology (NICT)), Japan from 1998 to 2010. In 2010, he moved to Tottori University, Japan, as a Professor in the Department of Information and Electronics, Graduate School of Engineering. His research interests include natural language processing, machine translation, and information retrieval.

Masato Tokuhisa

He received his Master's degree in Computer Science from Kyushu Institute of Technology, Japan, in 1995. He received his Doctor's degree in Engineering from Tottori University, Japan, in 2008. He was a research associate in Kyushu Institute of Technology from 1995 to 2002 and in Tottori University from 2002 to 2010. He is currently a junior associate professor in Tottori University.

Qing Ma

He received his B.S. degree in electrical engineering from Beijing University of Aeronautics and Astronautics (BUAA), China, in 1983, and the M.S. and Dr. Eng. degree in computer science from University of Tsukuba, Japan, in 1987 and 1990 respectively. He worked in Ono Sokki Co., Ltd., Japan from 1990 to 1993 and worked in the Communications Research Laboratory (Now: National Institute of Information and Communications Technology (NICT)), Japan from 1993 to 2003, as a Senior Researcher. In 2003, he moved to Ryukoku University, Japan, as a Professor in the Department of Applied Mathematics and Informatics, Faculty of Science and Technology. His research interests include machine learning and natural language processing.