

語句関連性とベイズ推論による対話型法律相談支援システム

MIZUKI, Ryutaro / 水木, 龍太郎

(出版者 / Publisher)

法政大学大学院情報科学研究科

(雑誌名 / Journal or Publication Title)

法政大学大学院紀要. 情報科学研究科編 / 法政大学大学院紀要. 情報科学研究科編

(巻 / Volume)

8

(開始ページ / Start Page)

207

(終了ページ / End Page)

210

(発行年 / Year)

2013-03

(URL)

<https://doi.org/10.15002/00009879>

語句関連性とベイズ推論による対話型法律相談支援システム

Words Correlation and Bayesian Inference based Interactive Legislation Consultation Support System

水木 龍太郎

Ryutaro Mizuki

法政大学大学院情報科学研究科情報科学専攻

E-mail: ryutaro.mizuki.5y@stu.hosei.ac.jp

Abstract

With increasing information and continuously updating information, it becomes difficult to find appropriate information needed. Traditionally, people consults with a lawyer when they are involved in a case related to a legal issue but it takes time and cost rather money. This paper describes a legal issue consultant support system which provides three functions: making graph expressed in the word relations; learning to know what a user wants to via the interactive process; and suggesting the related contents derived from Bayesian inference. Through interactions with the user, the system collects a sets of keywords related user needs, the system tries to find out what the user wants and retrieves them from the related graph database using Bayesian inference. Even for a beginner, it is easy to use since its operations are simple without requiring technical words. Moreover, the system is easy for administrator to manage. On the whole, the system is aimed at learning what a user needs, providing helpful information, and helping them to understand legislation.

1. 序論

法律は毎年新規追加・一部改正・廃止などがなされ、把握するには困難な勢いで変化していく。人々がある日唐突にトラブルに巻き込まれ自分の身を守るために弁護士と相談することは問題の解決には有効であるが、相談者に対して弁護士が不足していること、料金の問題など、一般人にとって弁護士に依頼するのは敷居が高い現状がある。また、インターネットで調べるといった方法は敷居が低いメリットがあるものの、相談者が予め法律関係の用語を知らないと言書の意味を理解できないというデメリットを持つ。それらの中間の解決策としてはインターネットでの相談という手段があるが、インターネット上の膨大な相談者に対して解決に必要な知識を持つ回答者は少なく、回答が返ってくる保障が無い割に回答者への負担が大きい。そこで本論文では相談者が求めている情報を絞り込み、法律に関する相談の解決を支援する対話型システムを提案する。システムでは、既存の法律相

談や法律に関するテキストに対してそれぞれに形態素解析を行い、特徴語とその依存関係を視覚化した語句関連グラフを文書毎に生成する。そして、相談者とシステム間の双方向の対話で関連している語句を推測し提案を繰り返して相談内容の語句関連グラフ化を行う。さらに、そのグラフに近似しているグラフを持つ文書をベイズ推定により推測し、ユーザーに相談内容と関連している文書を提案する。その結果、相談者は容易に相談を解決するための情報を手に入れることができ、円滑に問題を解決する手助けになる。

本システムでは対話に対話型遺伝的アルゴリズムが使われる。対話型遺伝的アルゴリズムとは遺伝的アルゴリズムの遺伝的操作を基にして、評価部分を人間の主観による提示個体の評価に置き換えたものである。解の評価を人間が行うため、進化の速度は低下し、世代交代数もあまり多く設定できなくなるが、対話型進化計算は発想支援システムとして捉えることができ、解が不完全な部分分解であってもそれをヒントに人間がアイデアを発想する、という可能性も考えられる。本論文では、対話型遺伝的アルゴリズムによって導かれたキーワードを用いて法律に関する文書を提案する。

2. 関連研究

2.1 電子政府総合窓口「e-Gov」

総務省行政管理局が運営している e-Gov 法令データ提供システム[1]は全法令をジャンル検索や 50 音順検索、一致検索で絞り込んで閲覧することができる。国営のため信頼性は高く無料で利用できるが、最低限の機能を持ったデータベースであり、ユーザーに法律の知識がないと利用するのは難しいと考えられる。本論で作成するシステムでは、この信頼性の高さを利用し、データを収集するサイトの 1 つとして採用している。

2.2 対話型遺伝的アルゴリズム

対話型遺伝的アルゴリズム[2]とは遺伝的アルゴリズムをベースとして、評価部分を人間の主観による個体の評価に置き換えたものである。人間の主観、感性で解の探索を行うため、定量的な評価が困難な楽曲やデザインの生成に適用されている。進化計算を用いて人間の主観的评价に基づく最適化手法を対話型進化計算と呼び、対話型遺伝的アルゴリズムは対話型進化計算の手法の 1 つである。

3. 法律相談支援システムの概要

本論文で扱うシステムは図1のように3つの要素によって構成されている。以下、実行が行われる順に記載する。

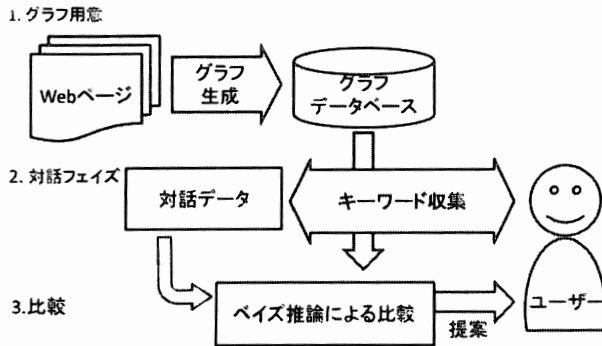


図1. 本システムにおける各要素の関係。

1) グラフ用意

まずシステムを稼働する前の準備として Web 上の法律相談や法律に関するウェブテキストを読み込み、一文毎に形態素解析で分解して、単語の文中における出現頻度から語句関連グラフを生成してデータベースに記録しておくことでユーザーへの提案に備える。

2) 対話フェイズ

ユーザーに年齢、職業、相談を1つの文章で入力してもらい、形態素解析で単語に分割する。そして、その各単語をデータベースに照合して合致した場合、その単語に関連するキーワードの一覧を作成してユーザーに提案して評価してもらい、要らないキーワードを順次交換していくことでキーワード群の更新を行い続ける。キーワード群の更新が行われなくなり次第、ベイズ推論フェイズへ移行する。

3) 比較

キーワード群からベイズ推論を用いて相談内容に合致する語句関連グラフを提案し、ユーザーの相談内容に役立ったのであれば良い評価としてデータベースにフィードバックすることでシステムの一連の流れとする。

これらの要素を組み合わせる構成したシステムの全体図が図2である。

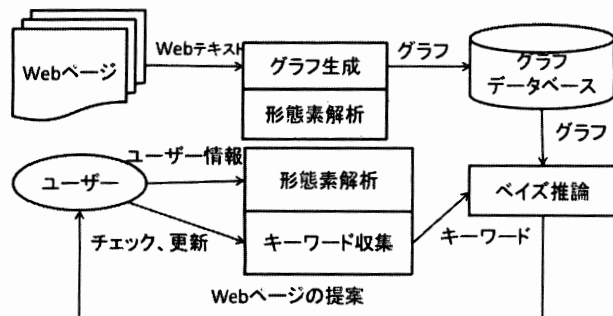


図2. 法律相談支援システム全体図。

3.1 形態素解析

本システムでは Web 上の法律相談文書及びユーザーが入力した文章を解析する手段として形態素解析を用いる。ユーザーから受け取った一言メッセージは形態素解析によって単語に分割されアルゴリズムに送られるが、形態素解析には Yahoo デベロッパーネットワークの WebAPI を使用した。Yahoo ならではの辞書の豊富さが強みであり、新しい単語にも対応しやすいためである。この API は、文章を受け取り、形態素解析した結果の単語、その単語の品詞、読み仮名のリストを出力する。本論では読み仮名は必要としていないので、単語と品詞のみを抜き出した後のグラフ生成やキーワード収集で利用している。

3.2 語句関連グラフ

本論文で提案するシステムは事前に Web 上から法律に関する文書や法律相談サイトでの相談文書を学習データとして収集し、それぞれの Web ページに形態素解析を用いて名詞や形容詞を抽出し、以下の定義に従って語句関連グラフを生成する。語句関連グラフとは、語彙連鎖グラフ(Word Colony)[2]を元に、本論文で取り扱う文書の時系列の要素を取り除き簡略化したものである。

1つの文書について語句関連グラフは複数のノードの繋がりと表され、1つのノードは1つの単語を表し、単語間の関連の度合いの強さを示す出現依存度とノードの文書における重要性を示す吸引力の2つの要素を持ち、図4のように表すことが出来る。出現依存度 $td(t, t')$ と吸引力 $attr(t)$ は以下の式で定義される。

$$td(t, t') = \frac{sentences(t \cap t')}{sentences(t)} \quad (1)$$

$$attr(t) = \sum_{t' \in T} td(t', t), \quad (t \neq t') \quad (2)$$

$sentences(t)$ は文書中における語 t が含まれる文の数であり、 $sentences(t \cap t')$ は t と t' を同時に含む文の数である。吸引力は語 t が文書中の他の語を引き付ける力とし、他の語から語 t に対する出現依存度の総和として定義する。

各語間の出現依存度を指標とすることにより、文書に出現する二つの語の間には

- 1) 双方向に強い依存度を持つ場合
- 2) 一方向にのみ強い依存を持つ場合
- 3) 双方向に弱い依存度を持つ場合

が考えられる。本システムではそれらに対して以下の行為を行う。

- 1) 双方向に強い依存度を持つ場合は2つの語の間には密接な関係が存在すると見做して1つのノードへの統合を行い、それらの吸引力も統合する。
- 2) 一方向にのみ強い依存を持つ場合は2つの語の間には依存するノードから依存されるノードへリンクを張る。
- 3) 双方向に弱い依存度を持つ場合は双方の間に出現に関する依存関係は無いと見做し、リンクは張らない。

そうすることで、文書の内容をキーワードとそれらの依存関係を持ったグラフとして表している。また、システムは全ての語句関連グラフを総合した独自の特徴語グラフを持ち、学習データ全体での語句の関連性データを持つ。例として法律相談サイトである弁護士ドットコムのある相談が掲載されているページをグラフ化する(図3)。このケースは個人情報の漏洩についての相談であり、「個人情報」や「メールアドレス」と言った単語が頻出している。「メールアドレス」や「住所」といった単語が出る文章には「個人情報」という単語があり、一方的な依存関係が存在したのでリンクが張られた。また、「住所」と「電話番号」は同じ文章に常に含まれたため、同一ノードにまとまり、吸引力が統合された(図4)。語句関連グラフを用いることによって、キーワードの関連性を表しつつ抽出できるため、キーワード提案の際に連想される用語を順に取り出すことが可能になる。

個人情報の漏洩について

13人が「役に立った」と評価 3回答済み 2弁護士回答 0お気に入り



谷田さん

会社のパソコンで顧客宛にダイレクトメールを送信した所、BCCにするのを忘れてしまい、多数のアドレスが「見える状態」で送信してしまいました。送信件数は120件ですが、1部の顧客から今後の対応を求められています。

見えてしまったのはアドレスのみで個人名・住所・電話番号等は一切含まれていません。アドレスのみでも個人情報に該当するのでしょうか？(フルネーム+企業名等の形式のアドレスはありませんでした)

図3. 学習用相談ケース。

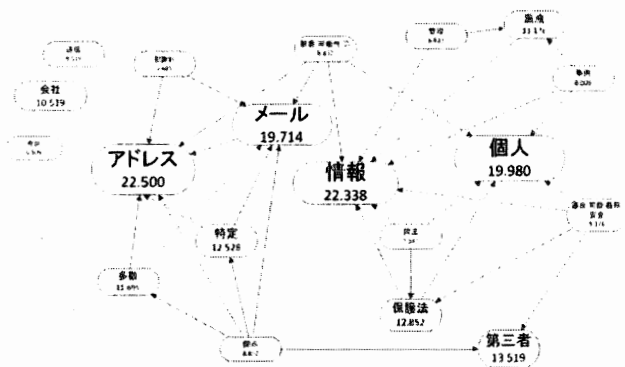


図4. 語句関連グラフ化の一例。

3.3 対話形式

序論でも触れたが、本論では対話部に対話型遺伝的アルゴリズム[2]を採用している。本システムではユーザーに職業、年齢、相談したい内容を1文で入力して貰い(図5)、文章に形態素解析を行った結果が初期個体に相当する。抽出される名詞からユーザーの相談内容に關係の有りそうなキーワードを特徴語グラフから複数探索して提案することが交叉に相当する。具体的には、キーワードを持ったノードを特徴語グラフから見つけ出し、そのノードにリンクが張られているノードを吸引力の高い順に選択

してキーワードを選出する。ユーザーは提案されたキーワード群から關係の有りそうなキーワードをできる限り選択しシステムに送信する。システムは選ばれたキーワードに關係の強いキーワードを選ばなかったキーワードの代わりに改めて用意し、また稀に突然変異として低確率でランダムに単語を選択したものを新しいキーワード群として生成して提案する。選択されたキーワードが十分な数が集まるか、ユーザーの評価が一定回数不変の場合に対話を終了し、終了するまでは新しいキーワードを用意してユーザーの評価を待ち続ける。例えばユーザーが「詐欺で悩んでいる」と入力した場合、「詐欺」に反応して「オークション」「ワンクリック」などと言った単語を提案し、「オークション」が選ばれたらまたそれに関連する「出品者」「落札者」といった単語が並ぶようになる。

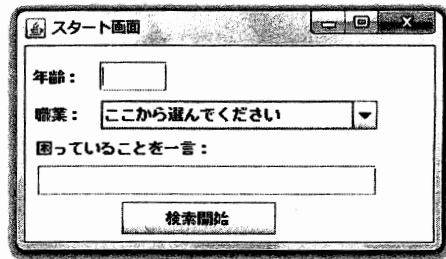


図5. ユーザー情報入力画面。

3.4 ベイズ推定による文書提案

まずユーザーから情報を入力された段階で選ばれたキーワード群から最尤法を用いてそれぞれの語句関連グラフが相談に合致する確率を計算する。大抵の場合、それだけで提案候補を決定するほどの確率にはならない。しかし、3.3で説明した対話を繰り返すことで必要と判断されたキーワードと不必要と判断されたキーワードが判明していく。そこでベイズ推定を用いて合致確率を再計算することによって提案候補の確率を対話毎に動的に変動させる。確率は次の計算式で求められる。

$$P(A|X) = \frac{P(X|A)P(A)}{\sum_B P(X|B)P(B)} \quad (3)$$

$P(A|X)$ は対話後の文書Aが提案される確率、 $P(A)$ は対話前の文書Aが提案される確率であり、 $P(X|A)$ は尤度である。一定の対話回数、もしくは合致確率が十分高い候補が現れた場合、その候補文書を提案する。その文書がユーザーを満足させられるものでは無かった場合はキーワードを再収集するが、キーワードが限界まで追加され新しく追加されなかった場合は合致確率が高く算出された順に提案を行う。例えば、「オークション」「詐欺」というキーワードだけではオークション詐欺の相談までしか絞れないが、「落札者」「送料」までのキーワードが判明すれば送料問題で悩んでいるということが判明し、より詳細に相談内容に近い文書を絞り込むことができる。

3.5 フィードバックによる強化学習

提案された文書が有用とユーザーに判断された場合、ユーザーの情報及び入力に使われた単語を提案された文書の語句関連グラフにユーザーの情報を語句関連グラフに追加する。既存単語は単語の出現回数に1を足し、新出単語は新規ノードとして追加する。それにより有用なキーワードは、よりユーザーに提案されやすくなるように学習される。

4. 実験

実際にあった事例「ナンパで何かしらの刑罰を科されることはあるのか?」[7]を用いて実験を行う。今回は、形態素解析で名詞のみを拾った場合と、名詞に加えて形容詞も拾った場合で語句関連グラフ化を行った。その結果が表1である。

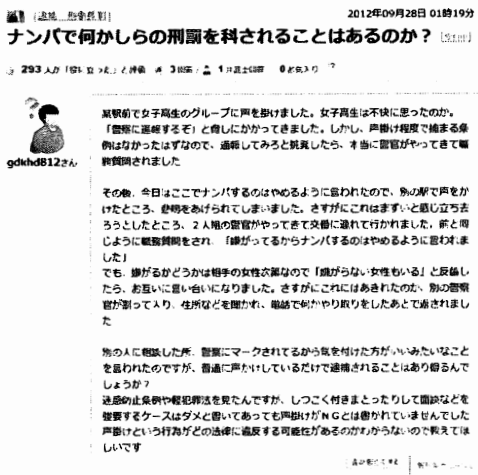


図6. 実験例[7].

表1. グラフ生成時の吸引力上位ノード.

名詞+形容詞	リンク先	吸引力	名詞のみ	リンク先	吸引力
声掛け		9.3	声掛け		6
声		7.8	条例	→声掛け	4.7
条例	→声掛け	7	職務質問		4.3
警官		6.8	通報		4.3
職務質問		5.3	警官		3.8
通報		5.3	警察		3.3
ナンパ		5.2	ナンパ		3
女子高生		5	声		3
警察		4.6	女子高生		2.5

グラフ生成結果、双方を比較しても上位ノードの項目は変わらず、あまりリンクが張られなかった。形容詞がもたらした効果として、形容詞自体は上位ノード、つまり重要単語にはならなかったものの、上位ノードに依存している事が判明した。例えば「しつこい」という単語は「警察」「声掛け」「条例」の単語に依存していることが確認された。また、表1の結果から「声」「警官」「女子高生」の3単語は形容詞の影響を受けて吸引力が大幅に増加したと考えられる。これらの単語は専門用語では無いことから、専門用語は形容詞に修飾さ

れにくい可能性を見出すことが出来た。また、この例は女子高生が相談者を不快に思っている、事実「不快」から「女子高生」にリンクが張られていたが、この結果からでは「相談者が女子高生を不快に思っている」のか「女子高生が相談者を不快に思っている」のか見分けることが出来ない。今回の場合は「女子高生が相談者を不快に思っている」のだが、これは「相談者」が文書中に出現しないために「相談者」の立ち位置を今回の解析では理解することが出来なかったと考えられる。

5. 考察

形態素解析で抽出する品詞は名詞のみとしたが、抽出する品詞を変更することでシステムの信頼度に十分影響を与える可能性がある。例えば名詞に加えて形容詞を抽出してキーワードとして使用した場合、個人の嗜好が表れやすくなるが、互いに相反するような感情が同じ文書に出現すると提案確率の上昇を阻害する可能性がある。また、実験結果から文書中に相談者が出てこないため、文書中の登場人物と相談者との関係が明確にならず、対人トラブルが発生した文書をうまく読み込めない欠点が発覚した。さらに、流行の相談に対応するためには学習データは頻繁に更新する必要があるが、学習データを取得する場所による偏りを無くすためにも大規模な運用が必要と成りうる。

6. 結論

本論文では、語句関連グラフやベイズ推定を用いて対話型法律相談支援システムを構築した。このシステムが実現化すれば、一般の人々にとってキーワードが分からなくてもより簡易な操作で相談に合う文書を提案することが出来る。規模の小さい相談であれば自己解決力の上昇も見込める。また、このシステムは学習型だが既に存在する文書を学習データとして利用できる所以他分野での利用も考えられる。

参考文献

- [1] e-Gov 法令データ提供システム, <http://www.e-gov.go.jp/>
- [2] 赤石 美奈, “文書群に対する物語構造の動的分解・再構成フレームワーク”, 人工知能学会論文誌, Vol.21, No.5 A(2006).
- [3] 井上 博行, 袁 丹, 岩谷 香栄, “対話型進化計算による配色作成支援システム”, 日本知能情報ファジィ学会誌, Vol.21, No.5 (2009).
- [4] 小野 智弘, 本村 陽一, 麻生 秀樹, “ベイズアンネットによる映画コンテンツの推薦方式の検討”, 電子情報通信学会技術報告, Vol.104, No.NC2004-66(2004).
- [5] 森 治憲, “ベイズ法における事前分布の持つ情報の評価法”, 日本統計学会誌, Vol.40, No.1(2010).
- [6] Bickel, Peter J. and Doksum, Kjell A. (2001). *Mathematical Statistics*, Vol.1, No.2 (2007) ISBN 0-13-850363-X.
- [7] ナンパで何かしらの刑罰を科されることはあるのか? 一弁護士ドットコム <http://www.bengo4.com/bbs/141803>