Studia Iuridica 83
ISSN 0137-4346, e-ISSN 2544-3135

DOI: https://doi.org/10.31338/2544-3135.si.2020-83.12

Jacek Petzel
Uniwersytet Warszawski, Polska
ORCID: 0000-0001-5728-8373

STATYSTYCZNE WŁAŚCIWOŚCI TEKSTÓW PRAWNYCH I ICH WYKORZYSTANIE W SYSTEMACH WYSZUKIWANIA INFORMACJI PRAWNEJ

Współczesne systemy wyszukiwania informacji prawnej w olbrzymiej większości opierają się na wykorzystaniu w operacjach wyszukiwawczych strategii pełnego tekstu (*full text*). Polega ona na tym, iż po wprowadzeniu do urządzeń pamięciowych komputera pełnych tekstów dokumentów i wygenerowaniu przez komputer zapisów służących do przeszukania bazy danych, użytkownik systemu może odnaleźć określony dokument poprzez użycie któregokolwiek ze słów występujących w tekście tego dokumentu. Ta metoda wyszukiwania, która notabene stosowana jest już od prawie 60 lat, bowiem pierwszy system o tym charakterze został stworzony w roku 1960 przez J. Horty'ego, już w latach 80. ubiegłego stulecia poddawana była krytyce. W jej ramach podnoszono w szczególności, że wyszukiwanie oparte na słowach, nawet w sytuacji, kiedy istnieje możliwość formułowania instrukcji wyszukiwawczych wykorzystujących operatory logiczne, a także pozycyjne, nie zaspokaja w dostatecznym stopniu potrzeb użytkowników systemów[1]. Pojawiły się wówczas pierwsze koncepcje tworzenia systemów, w których operacja wyszukiwania opierałaby się na innych, nowocześniejszych rozwiązaniach. Systemy te określamy jako systemy konceptualne, choć niekoniecznie wiążą się one z przeprowadzaniem wyszukiwania na podstawie pojęcia prawnego (*concept*). Nazwą tą określa się obecnie ogół różnorodnych metod proponowanych w informatyce prawniczej dla wyszukiwania informacji,

[1] Zauważył to, co ciekawe, już twórca systemów pełnotekstowych J. Horty. Na ten temat patrz: J. Bing, *Legal Decisions and Information Systems*, Universitetsforlaget, Oslo 1977, s. 63.

które starają się stosować rozwiązania bardziej nowoczesne niż rozwiązania pełnotekstowe.

Niniejszy artykuł będzie poświęcony zagadnieniom wiążącym się z problematyką wykorzystywania dla celu wyszukiwania informacji prawnej właściwości statystycznych tekstów prawnych, jakie są zawarte w bazach dokumentacyjnych systemów wyszukiwawczych. W szczególności będą one dotyczyć omówienia metod pozwalających na rozszerzenie zbioru dokumentów wyszukanych o dokumenty semantycznie tym dokumentom bliskie, tj. dotyczące tej samej materii treściowej. Zdaniem wielu specjalistów przedmiotowych ten sposób przeprowadzenia wyszukania pozwala na zaspokojenie w lepszym stopniu potrzeb informacyjnych użytkowników systemu. Co więcej, przy odpowiednim przeprowadzeniu tej operacji realizować może główne założenie wiążące się z wyszukiwaniem informacji, w tym informacji prawnej, które sprowadza się do stwierdzenia, że w wyniku jej przeprowadzenia użytkownik systemu powinien otrzymać nie tylko te dokumenty, których szuka, ale również te, które są mu obiektywnie potrzebne dla zaspokojenia jego potrzeb informacyjnych.

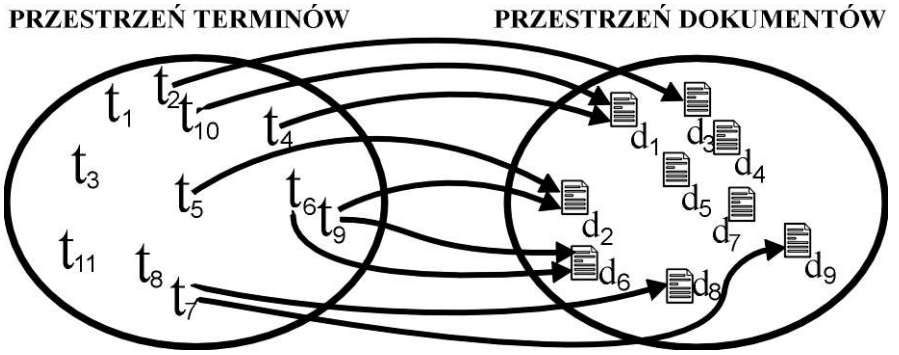
Osiągnięcie takiego celu jest możliwe poprzez dokonywanie zabiegów zarówno w „przestrzeni semantycznej terminów", jak i „przestrzeni semantycznej dokumentów". Przez termin „przestrzeń semantyczna terminów" będziemy rozumieć taką n-wymiarową przestrzeń, w której umieszczonych jest n-terminów wchodzących w skład dokumentów stanowiących bazę dokumentacyjną systemu. Liczba wymiarów w tej przestrzeni jest równa liczbie terminów znajdujących się w dokumentach. W przestrzeni tej znajdują się wszystkie słowa (określone jako *word type*) występujące w tekstach bazy dokumentacyjnej z wyjątkiem słów uznanych za nieużyteczne dla indeksowania. Położenie każdego terminu reprezentowanego przez określony punkt w tej przestrzeni nie jest przypadkowe. Leżą one w pewnej bliskości lub oddaleniu od siebie w zależności od tego, jakie materie treściowe wyrażają. Dla przykładu termin „kara" znajduje się w tej przestrzeni bliżej terminu „zabija" niż termin „małżeństwo". Jest on bowiem silniej powiązany semantycznie z tym pierwszym niż z tym drugim.

Przez termin „przestrzeń semantyczna dokumentów" będziemy rozumieć z kolei taką n-wymiarową przestrzeń, w której umieszczone są wszystkie dokumenty składające się na bazę dokumentacyjną systemu. Liczba wymiarów tej przestrzeni jest równa liczbie znajdujących się w niej jednostek tekstowych. Każdy z dokumentów reprezentowany jest przez pewien punkt w tej przestrzeni. Położenie tych punktów nie jest przypadkowe i jest uzależnione od tego, jak bliskie sobie problemy poruszają reprezentowane przez nie teksty. Dla przykładu punkt reprezentujący dokument będący art. 31 Kodeksu rodzinnego i opiekuńczego (tekst jedn. Dz.U. z 2019 r., poz. 2086, z późn. zm.) będzie leżał w tej przestrzeni bliżej punktu reprezentującego dokument będący art. 42 tego kodeksu niż punktu odtwarzającego art. 11. Artykuły 31 i 42 regulują bowiem kwestie związane ze wspólnotą ustawową małżonków, a więc bliskie sobie materie treściowe,

zaś art. 11 mówiący o wpływie ubezwłasnowolnienia na możliwość zawarcia związku małżeńskiego dotyczy innej problematyki.

Poniżej (rys. 1) przedstawiona jest „przestrzeń terminów", „przestrzeń dokumentów" oraz występujące między nimi powiązania. Ze względu na przejrzystość rysunku oznaczono jedynie niewielką część sieci powiązań. Ponieważ każdy dokument składa się z wielu identyfikujących go terminów, a jeden termin może identyfikować wiele dokumentów, rysunek ukazujący wszystkie możliwe powiązania byłby nieczytelny.

Rys. 1. Powiązania między „przestrzenią terminów" a „przestrzenią dokumentów"



Źródło: Opracowanie własne

Osiągnięcie rezultatu polegającego na tym, że system w wyniku przeprowadzenia operacji wyszukiwania odnajdzie nie tylko te dokumenty, w których występują słowa, jakich użytkownik użył w instrukcji, lecz także inne bliskie semantycznie tym dokumentom, może być realizowany w dwojaki sposób. Po pierwsze, poprzez tworzenie podprzestrzeni w przestrzeni terminów. Jej stworzenie jest możliwe poprzez wykorzystanie dla mierzenia w tej przestrzeni odległości między terminami za pomocą miar mierzących siłę korelacji określonych terminów w bazie dokumentacyjnej. Przyjmuje się przy tym założenie, że im silniej określone terminy korelują ze sobą w tekstach, tym są sobie bliższe znaczeniowo. Informacje o sile korelacji terminów są zawarte w tezaurusie statystycznym, a ich wyszukiwanie pozwala na rozszerzenie instrukcji wyszukiwawczej o terminy semantycznie bliskie terminom użytym przez użytkownika przy przyjęciu założenia, że silne korelowanie ze sobą terminów w tekstach świadczy o ich bliskości znaczeniowej. Zakłada się przy tym, iż bliskie semantycznie terminy będą wyszukiwać bliskie semantycznie dokumenty. W wyniku przeprowadzenia operacji użytkownik systemu ma otrzymywać zbiór bliskich semantycznie dokumentów, określonych jako wiązka (*cluster*). Po drugie, poprzez tworzenie podprzestrzeni w przestrzeni dokumentów, co jest możliwe do zrealizowania za pomocą stosowania różnych metod, wśród których za najistotniejsze należy uznać metodę analizy wektorowej, metody wykorzystania sieci neurono-

wych, a także wiele innych, w tym będącą przedmiotem naszego zainteresowania metodę statystycznej analizy podobieństwa tekstów. Analizy tej dokonuje się za pomocą wykorzystania specjalnie stworzonych miar (współczynników korelacji) mających pozwalać na obliczenie stopnia, w jakim dokumenty są do siebie podobne. W rezultacie tej operacji, również i w tej sytuacji użytkownik ma otrzymywać wiązkę dokumentów (*cluster*), co do których zakłada się, że są sobie bliskie semantycznie.

Należy zauważyć, że stosowanie zarówno pierwszej, jak i drugiej metody opiera się na założeniu o odpowiedniości, które zakłada, że istnieje zależność pomiędzy właściwościami statystycznymi tekstów a ich właściwościami semantycznymi. Ma ono charakter idealistyczny, ale przyjmuje się, że związek ten występuje w dostatecznej liczbie przypadków, aby było możliwe wykorzystanie go w praktyce. O tym, że zależność taka ma miejsce, świadczy to, że z występowania w ramach dwóch dokumentów tych samych słów z dużą dozą prawdopodobieństwa można wysnuć wniosek, iż dotyczą one tej samej bądź bardzo podobnej materii treściowej.

Idea badania podobieństwa dokumentów oparta na wykorzystaniu właściwości tekstów – choć, mówiąc precyzyjniej, chodziło tu raczej o wykorzystanie indeksów przypisanych dokumentom – po raz pierwszy została zastosowana przez G. Saltona w systemie SMART już w 1971 r. Podobieństwo dokumentów zostało określone przez niego jako stopień, w którym dwa dokumenty są do siebie zbliżone, ze względu na formę, w szczególności na użycie takich samych indeksów dla opisu ich zawartości treściowej. Ponieważ system SMART był systemem indeksowym, podobieństwa tego nie dało się liczyć w sposób automatyczny przy wykorzystaniu statystyki tekstu. W 1981 r. A. F. Smeaton i C. J. van Rijsbergen[2] wprowadzili koncepcję dokumentu bazowego, uznając, że za podlegające wyszukaniu powinny być uznane wszystkie te dokumenty, w których tekście znajdują się w odpowiednim nasileniu te same słowa co w dokumencie bazowym. Podobna metoda była proponowana przez P. Wiletta[3]. V. Batagelj i M. Bren[4] w 1993 r. Dokonując przeglądu stosowanych dla badania podobieństwa dokumentów dwudziestu dwóch miar, wskazali, że dla określenia zbioru dokumentów wyszukanych istotne jest nie tylko badanie stopnia podobieństwa dokumentów, ale również stopnia braku ich podobieństwa. Cechą wszystkich tych metod było to, że nie przypisywało się różnych wag do poszczególnych słów (*word types*).

Proponowane współcześnie na gruncie informatyki prawniczej metody badania podobieństwa tekstów wiążą się z nowymi możliwościami, które otwiera przed

[2] A. F. Smeaton, C. J. van Rijsbergen, *The Nearest Neighbour Problem in Information Retrieval*, Proceedings of the Fourth International Conference on Information Storage and Retrieval, ACM, New York 1981, s. 83–87.

[3] P. Willet, *A Fast Procedure for the Calcuulation of Similarity Coefficients in Autmoatic Classification*, (w:) *Information Processing and Management*, Vol. 17, Oxfor 1981, s. 53–60.

[4] V. Batagelj, M. Bren, *Comparing Similarity Measures*, Ljubljana 1993.

badaczami znaczne zwiększenie pamięci operacyjnej komputerów, co umożliwia określanie wzajemnego podobieństwa dokumentów nawet w bardzo obszernych ich zbiorach[5]. Jednymi z istotnych badań, w których wykorzystywano miary statystyczne dla badania podobieństwa dokumentów, były badania prowadzone przez K. van Noortwijka i R. V. De Muldera w ramach programu ERASMUS na Uniwersytecie w Rotterdamie[6].

Badając podobieństwa dokumentów, K. van Noortwijk i R. V de Mulder wskazali, iż dokumenty mogą być podobne ze względu na różne kryteria, do których należą: 1) użyta czcionka, 2) ogólny wygląd – format stron czy też nagłówków, 3) liczba słów występujących w dokumencie i liczba różnych słów występujących w dokumencie, 4) frekwencja występowania różnych słów w dokumencie, 5) długość słów, zdań, paragrafów lub stron, 6) język dokumentów, 7) adresat dokumentów, 8) przedmiot, którego dotyczą, 9) efekt, jaki powodują ze względu na czas publikacji bądź czas zapoznania się z ich treścią. Podobieństwa pkt 1–5 mają charakter podobieństwa czysto formalnego, pkt 6–7 mają charakter mieszany, podobieństwo 8 ma charakter semantyczny, a pkt 9 – pragmatyczny. Uznali oni przy tym, że najistotniejsze z podobieństw, jakie można wykorzystać w celu badania podobieństwa dokumentów, to frekwencja występowania różnych słów w różnych dokumentach oraz liczba różnych słów występujących w jednym dokumencie.

Podstawowymi parametrami, jakie biorą oni pod uwagę przy obliczaniu stopnia podobieństwa dokumentów, są współwystępowanie, a także przy stosowaniu niektórych współczynników korelacji niewspółwystępowanie ze sobą określonych słów w tekstach porównywanych dokumentów. Wskazują przy tym na różne rodzaje współwystępowania i niewspółwystępowania. Po pierwsze ich zdaniem może być tak, że pewne słowo występuje w obu dokumentach, co wskazuje na ich podobieństwo. Sytuację taką określają jako *hit*. Wyodrębniają jednak dwa rodzaje *hit*. *Hit* (1) typu pierwszego polega na tym, że słowo obecne jest w obu dokumentach, i *hit* (2) polegający na tym, że słowa nie ma w obu dokumentach, ale występują w innych dokumentach bazy. Oba typy *hitów* uznają za wskazujące na występowanie podobieństwa dokumentów, co, jak się wydaje, w przypadku *hit* (2) jest dyskusyjne. Z kolei jeśli określone słowo występuje w jednym dokumencie, a nie występuje w drugim, to określają to jako *miss*. Wyodrębniają jednak również dwa rodzaje *miss*, co związane jest z tym, że wartość przyjmowana przez

[5] Ze względu na małą pojemność pamięci operacyjnej w przeszłości nie było możliwości ani stosowania metod analizy wektorowej, ani też metod opartych na statystycznych miarach korelacji dla celów wyodrębniania podzbiorów w przestrzeni dokumentów. Stąd też badania koncentrowały się na tworzeniu tych podprzestrzeni w zbiorach terminów w oparciu o tezaurusy statystyczne. Przyjmowano przy tym założenie, że silnie korelujące ze sobą w tekstach terminy, co świadczyć miało o ich podobieństwie semantycznym, identyfikować będą podobne semantycznie dokumenty.

[6] K. van Noortwijk, R. V. De Mulder, *The Similarities of Text Documents*, "The Journal of Information, Law and Technology" 1997, Vol. 2.

miss jest uzależniona od tego, jaki dokument jest uznawany za dokument bazowy, co jest istotne zwłaszcza w sytuacji, w której dokumenty nie są równe ze względu o swoją wielkość. Dla przykładu, jeżeli porównuje się dwa dokumenty, z których pierwszy A zawiera 100 *word types* (różnych słów), a drugi B 200 takich słów, a jednocześnie w obu dokumentach znajduje się 50 takich samych *word types*, to w sytuacji kiedy porównamy dokument A zawierający 100 różnych słów z dokumentem B, to *misses* stanowią 50 słów. Natomiast jeżeli porównamy ze sobą dokument B zawierający 200 słów z dokumentem A, to *misses* będą stanowić 150 słów. Oznacza to, że badanie wartości określanej przez *misses* ma charakter relatywny i zależy od tego, który dokument przyjmiemy jako bazowy. W konsekwencji ich zdaniem nie tylko musimy porównywać dokument A z B, ale także B z A. Tego typu rozróżnienie występowało już uprzednio w literaturze opisującej różnorodne współczynniki korelacji, przy czym *hits* i *misses* oznaczano literami: *hit* (1) – a, *miss* (1) – b, *miss* (2) – c, *hit* (2) – d[7].

Tabela 1. Współczynniki korelacji

	A	**~ A**	
B	**a (hit 1)**	**b (miss 1)**	**a+b**
~B	**c (miss 2)**	**d (hit 2)**	**c+d**
	a+c	b+d	N

Źródło: Opracowaie własne.

Twórcy koncepcji badania podobieństwa dokumentów zastosowanej w programie ERASMUS zwrócili jednak uwagę na coś, co dotychczas nie było elementem uwzględnianym przy obliczaniu współczynników podobieństwa, a mianowicie na to, że każdemu *hitowi*, a także *missowi* może być przypisana określona waga. Konieczność przypisywania określonym *hitowi* i *miss* wag wynika ich zdaniem z tego, że w zależności od częstości użycia tych słów w dokumentach bazy dokumentacyjnej, wartość *hit* (1) i *hit* (2) powinna brać pod uwagę fakt, że w sytuacji słów, które bardzo często występują w ramach bazy dokumentacyjnej, prawdopodobieństwo, że pewne słowa będą występować w obu dokumentach, jest dużo większe niż w sytuacji, w której słowa te są słowami rzadko występującymi w dokumentach. Podobnie przedstawia się sytuacja, jeśli chodzi o *misses*. Uznali więc, że konieczne jest obliczenie prawdopodobieństwa wystąpienia określonego słowa w ramach dokumentów, co jest istotne dla określenia realnej wartości *hit* (1) i *miss* (2), oraz prawdopodobieństwa braku określonego słowa w określonym dokumencie, co jest istotne dla określenia realnej wartości *hit* (2) i *miss* (1). W celu obliczenia tego prawdopodobieństwa proponują oni dwie miary, gdzie F_i oznacza ilość typów słów występujących w bazie dokumentacyjnej, zaś D liczbę dokumentów w całej bazie dokumentacyjnej.

[7] Patrz Z. Rogoziński, *Metody statystyczne w prawoznawstwie*, Warszawa 1976, s. 142.

$$P_{(1hit1,miss2)} = \frac{F_i}{D} \qquad P_{(1hit2,miss2)} = \frac{D - F_i}{D}$$

Używając tych dwóch parametrów, proponują oni zastosowanie dla obliczenia wagi poszczególnych wartości *hit* i *misses* dla każdego z *word type.* Współczynnikiem, który ma określać tę wartość, jest:

$$W_{(i)} = 1 - P_{(i)}$$

Zastosowanie wag, w ich przekonaniu, pozwoliło na określenie tego, co uznają za lepiej oddającą stan rzeczywisty wartość parametrów *hit* i *miss*. Wskazali przy tym słusznie, że prawdopodobieństwo wystąpienia *hit* (1) w sytuacji, gdy określone słowo występuje w wielu dokumentach, jest większe, ale znaczenie takiego *hit* (1) dla określenia podobieństwa jest niskie. Oznacza to ich zdaniem konieczność przemnożenia każdej z wartości osiąganych przez *hit* i *miss* przez współczynnik W, którego wartość jest odwrotna w stosunku do prawdopodobieństwa wystąpienia słowa. Użycie wag pozwala na obliczenie dla każdego *word type* skorygowanych wartości *hits* i *misses*. Wartość ta jest równa sumie wszystkich wag osiąganych przez każde ze słów. Dla przykładu, skorygowana wartość *hit* (1) jest równa:

$$Hit1_{adj} = \sum_{i=1}^{n} W(i)$$

Tak obliczona wartość w przypadku porównywania podobieństwa pary dokumentów musi być ich zdaniem w dalszej kolejności zrelatywizowana do maksymalnych wartości, jaką mogą one osiągnąć. Wartości te są różne dla różnych dokumentów. Maksymalna wartość w przypadku *hit* (1) i *miss* (1) jest sumą wag występujących w dokumencie, a w przypadku *hit* (2) i *miss* (2) sumą wagą słów niewystępujących w dokumencie. Po podzieleniu przypisanych określonym *word types hit* i *misses* przez ich wartości maksymalne i przemnożeniu ich przez 100% otrzymujemy zdaniem realizujących program ERASMUS rzeczywistą wartość procentową *hits* i *misses*, które umożliwiają porównanie par dokumentów w bazie. Niezależnie od tego, przy obliczaniu podobieństwa dokumentów, przy stosowaniu niektórych testowanych współczynników bierze się również pod uwagę miarę wskazującą na ilość informacji, która pozwala na przypisanie większej wagi słowom rzadko występującym w bazie dokumentacyjnej. Słowa takie niosą bowiem więcej informacji niż słowa często w bazie występujące. Dla tego celu używają oni miary:

$$L_{(a)}^{2} = \log \frac{1}{P_{(a)}}$$

Następnie proponują oni stosowanie trzech różnych współczynników badania podobieństwa, z których za najbardziej przydatny uznają[8]:

$$S = \frac{\text{Hit1adj} + \text{Hit2adj}}{\text{Hit1max} + \text{Hit2max}}$$

Wskazują jednak, że w niektórych sytuacjach bardziej właściwe jest stosowanie innych współczynników[9]. Wskazują również na dwa współczynniki używające zamiast wag ilości informacji, jaką niosą ze sobą poszczególne słowa występujące w dokumentach. Jednym z takich współczynników jest:

$$S = \frac{\text{Hit1info}}{\text{Hit1maxinfo}}$$

Procedura badania podobieństwa dokumentów polega na stosowaniu kolejnych kroków. Po pierwsze po przeprowadzeniu operacji wyszukania w sposób klasyczny użytkownik wybiera spośród wyszukanych dokumentów tzw. dokument wzorcowy, za który uznawany jest ten dokument, który zdaniem użytkownika jest dokumentem najbardziej relewantnym. Następnie dokonuje się porównania tego dokumentu z innymi dokumentami przy wykorzystaniu listy różnych słów (*word types*), jakie w nich występują, listy różnych słów występujących w bazie dokumentacyjnej, a także listy słów wskazujących na liczbę dokumentów, w których dane słowo występuje, co pozwala na określenie wartości *hits* i *misses*, a w przypadku wskazanego przez nich jako najbardziej przydatny współczynnika jedynie wartości *hit* (1), *hit* (2), które koryguje się za pomocą wykorzystania wag, uzyskując Hit1adj i Hit2adj. Na kolejnym etapie oblicza się ich wartości maksymalne Hit1max i Hit2max. W rezultacie przy wykorzystaniu wartości współczynnika korelowania poszczególnych słów w tekstach porównywanych dokumentów sporządzana jest uporządkowana lista dokumentów biorąca pod uwagę stopień podobieństwa każdego z nich do dokumentu wzorca[10]. Na początku tej listy (*ranked at the top*) powinny znajdować się dokumenty, które są najbardziej relewantnymi. Jeżeli wśród nich znajdują się dokumenty, które użytkownik uzna za wysoce relewantne, może dodać je do listy wzorców. W konsekwencji dojdzie do ponownego przeszukania bazy danych, opierając się na słowach zawartych w nowych wzorcach. Spowoduje to powstanie nowej listy dokumentów wyszukanych. Może

[8] K. van Noortwijk i R. V. de Mulder, proponując stosowanie skonstruowanych przez siebie współczynników, podkreślili, iż ich zdaniem żaden z opisanych przez V. Batagelja i M. Brena nie powinien być stosowany.

[9] Na temat tych współczynników patrz bliżej K. van Noortwijk, R. V. de Mulder, *The Similarities...*

[10] Twórcy systemu badali również możliwość stosowania innych współczynników korelacji, przy czym mimo wyodrębnienia różnych kategorii *misses* brali je pod uwagę tylko przy badaniu jednego współczynnika. Współczynnikiem tym jest $S = \frac{\text{Hit1adj} + \text{Hit2adj} - \text{Miss1adj} - \text{Miss2adj}}{\text{Hit1max} + \text{Hit2max}}$. Uznali więc słusznie, iż chodzi o badanie podobieństwa, a nie niepodobieństwa dokumentów.

się jednak zdarzyć, że wśród dokumentów przedstawionych użytkownikowi na pierwszym etapie znajdują się dokumenty, które mimo iż formalnie określane są jako wysoce relewantne, w rzeczywistości takimi nie są. Dokumenty takie powinny być wskazane przez użytkownika i dostarczone do systemu jako dokumenty kontrawzorce. Dokumenty te charakteryzuje to, że obliczona za pomocą kryteriów formalnych siła ich podobieństwa jest wysoka, a mimo to nie są w rzeczywistości relewantnymi[11]. Wskazanie przez użytkownika systemu nowych dokumentów kontrawzorców powoduje ponowne przeszukanie bazy dokumentacyjnej. Komputer jeszcze raz oblicza stopień podobieństwa dokumentów i prezentuje nową uporządkowaną ich listę. W takiej sytuacji może się zdarzyć, że nowo dostarczona lista dokumentów zawiera nowe wzorce zawierające słowa, które nie były prezentowane w poprzednich wzorcach, co może prowadzić do przeprowadzenia nowego przeszukania bazy danych. Mogą również pojawić się nowe dokumenty będące kontrawzorcami. Prowadzić to będzie do kolejnych przeszukiwań bazy danych, a cała operacja, mająca charakter iteracyjny, może być kontynuowana aż do momentu, w którym użytkownik stwierdzi, że kolejne wyszukiwanie nie prowadzi już do istotnej zmiany pozycji dokumentu na liście rankingowej. Proponowana przez twórców tej koncepcji metoda była testowana na stosunkowo dużej bazie zawierającej 18 803 dokumenty, w której występowało 143 156 różnych słów. Brak jest informacji o tym, jakie były rezultaty tego testowania, co pozwala domniemywać, że nie były one zbyt dobre. Powodem, dla którego osiągnięty rezultat nie był w pełni zadowalający – że takim był, świadczy choćby to, iż system ten nie jest w praktyce wykorzystywany do wyszukiwania informacji – było, jak się wydaje, niedokonanie analizy semantycznej dokumentów wprowadzanych do bazy. Ograniczano się bowiem, prawdopodobnie, do określenia *word types*, bowiem brak jest informacji o tym, że uwzględniano zjawiska synonimiczności czy też homonimiczności, jakie zawsze zachodzą między słowami w dokumentach składających się na bazę dokumentacyjną. Nie oznacza to oczywiście, że metoda polegająca na badaniu podobieństwa dokumentów nie jest właściwa dla tworzenia wiązek semantycznie bliskich dokumentów. Osiągnięcie właściwych rezultatów jest jednak uzależnione od wielu czynników, takich jak dobór odpowiedniego współczynnika korelacji, określenie tego, w jaki sposób ma być realizowana operacja wyodrębniania podprzestrzeni dokumentów w oparciu o dokumenty wzorce (dokumenty bazowe), a także od tego, jaki jest poziom analizy semantycznej tekstów wprowadzanych do bazy dokumentacyjnej. Stąd też badanie użyteczności tej metody wymaga kontynuowania prac

[11] Stwierdzenie występowania relewancji określonych dokumentów prawnych jest procesem skomplikowanym. Patrz bliżej F. Studnicki, *Kryterium relewancji w zautomatyzowanym wyszukiwaniu informacji prawnej*, (w:) *Prawne problemy systemów informatycznych. Konferencja informatyki prawniczej*, Wrocław 1976, t. 2, s. 23–39. Patrz również J. Petzel, *Relewancja tekstów prawnych*, (w:) *Prawo handlowe XXI wieku. Czas stabilizacji, ewolucji czy rewolucji. Księga jubileuszowa Profesora Józefa Okolskiego*, Warszawa 2010, s. 767–778.

badawczych. Należy jednak przy tym pamiętać, że w opinii wielu badaczy *no one algorithm will give you all relevant results*[12].

BIBLIOGRAFIA

Batagelj V., Bren M., *Comparing Similarity Measures*, Ljubljana 1993

Bing J., *Legal Decisions and Information Systems*, Universitetsforlaget, Oslo 1977

Mart S. N., *The Relevance of Results Generated by Human Indexing and Computer Algorithms: A Study of West's Headnotes and Key Numbers and LexisNexis's Headnotes and Topics*, "Law Library Journal" 2010, Vol. 102

Noortwijk K. van, De Mulder R. V., *The Similarities of Text Documents*, "The Journal of Information, Law and Technology" 1997, Vol. 2

Petzel J., *Informatyka prawnicza*, *zagadnienia teorii I praktyki*, Warszawa 1999

Petzel J., *Relewancja tekstów prawnych*, (w:) *Prawo handlowe XXI wieku. Czas stabilizacji, ewolucji czy rewolucji. Księga jubileuszowa Profesora Józefa Okolskiego*, Warszawa 2010

Rogoziński Z., *Metody statystyczne w prawoznawstwie*, Warszawa 1976

Smeaton A. F., van Rijsbergen C. J., *The Nearest Neighbour Problem in Information Retrieval*, Proceedings of the Fourth International Conference on Information Storage and Retrieval, ACM, New York 1981

Studnicki F., *Kryterium relewancji w zautomatyzowanym wyszukiwaniu informacji prawnej*, (w:) *Prawne problemy systemów informatycznych. Konferencja informatyki prawniczej*, t. 2, Wrocław 1976

Willet P., *A Fast Procedure for the Calculation of Similarity Coefficients in Autmoatic Classification*, (w:) *Information Processing and Management*, Vol. 17, Oxford 1981

STATISTICAL PROPERTIES OF LEGAL TEXTS AND THEIR USE IN LEGAL INFORMATION RETRIEVAL SYSTEMS

Summary

The article is devoted to issues related to the use of statistical properties of legal texts for searching legal information. Methods are presented that allow to enlarge the set of found documents by including those semantically close to the initially found ones. Providing the ability to search for such collections of documents allows to better satisfy the

[12] Patrz S. N. Mart, *The Relevance of Results Generated by Human Indexing and Computer Algorithms: A Study of West's Headnotes and Key Numbers and LexisNexis's Headnotes and Topics*, "Law Library Journal" 2010, Vol. 102, s. 13.

needs of the system users. The article presents the theoretical foundations of enlargement operations based on performing specific treatments in the so-called *semantic space of terms* and *semantic space of document*s. The article deals particularly with the methods which allow to determine the set of similar documents by using statistical properties of documents. The research carried out under the ERASMUS program conducted by K. van Noortwijk and R. V. de Mulder is presented in detail. A critical analysis of the measures used in this research was carried out, as well as an analysis of the reasons why the proposed methods didn't lead to fully satisfactory results.

KEYWORDS

legal informatics, statistical properties of legal texts, measures of correlation, similarity of documents in legal databases, ERASMUS program

SŁOWA KLUCZOWE

informatyka prawnicza, statystyczne właściwości tekstów prawnych, miary korelacji, podobieństwo dokumentów, program ERASMUS