PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
Nr 7 (1207) 2008
TAKSONOMIA 15
Klasyfikacja i analiza danych – teoria i zastosowania

Mariusz Grabowski
Uniwersytet Ekonomiczny w Krakowie

WYKORZYSTANIE METOD EKSPLORACYJNEJ ANALIZY TEKSTU DO IDENTYFIKACJI KLUCZOWYCH ZAGADNIEŃ ZAWARTYCH W DUŻYCH ZBIORACH PUBLIKACJI NAUKOWYCH

1. Wstęp

Obserwowany w ostatniej dekadzie rozwój metod i środków przechowywania i udostępniania danych zaowocował zwiększeniem dostępności zbiorów informacji, a szczególnie dokumentów tekstowych. Prognozowany w pracy [Ackoff 1967] zalew informacji stał się faktem, a postulowana funkcjonalność systemów informacyjnych, polegająca na zadaniach selekcji i kondensacji informacji – koniecznością. Rozwój metod analizy danych, zwłaszcza narzędzi należących do ogólnej kategorii eksploracyjnej analizy tekstu (EAT) (*text mining*) pozwala w wielu przypadkach na automatyczną realizację pierwszej z wyżej wymienionych funkcjonalności.

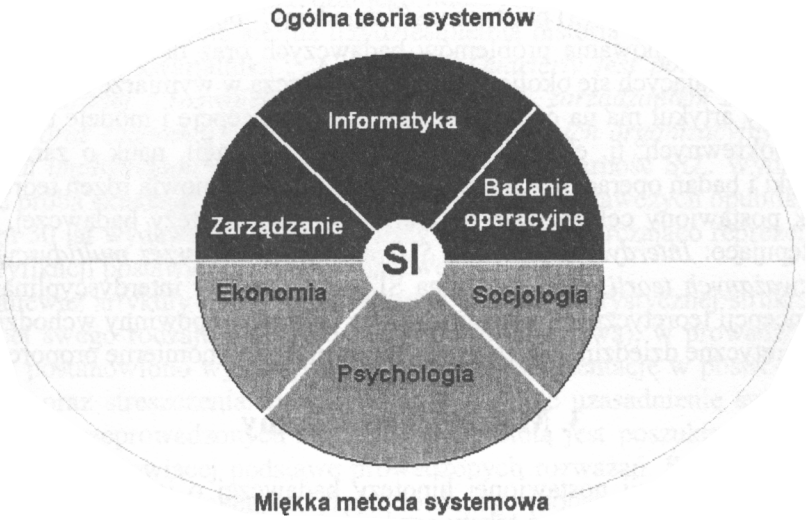
Pomysł użycia metod i środków informatyki do analizy dokumentów tekstowych ma już dość długą historię. Należy tutaj przede wszystkim wymienić prace Chomsky'ego, test Turinga czy opracowanie przez Weizenbauma programu *Eliza*. Podejściem znajdującym wiele zastosowań praktycznych jest ostatnio właśnie eksploracyjna analiza tekstu [Hearst 1999]. Eksploracyjna analiza tekstu wykorzystuje metody drążenia danych [Witten, Frank 2000] w odniesieniu do dokumentów tekstowych. Jedna z pierwszej definicji eksploracyjnej analizy tekstu została zaproponowana w pracy [Hearst 2003], według której jest nią „odkrywanie przez komputer nowych, wcześniej nieznanych informacji, dzięki automatycznemu wydobywaniu informacji z różnych źródeł tekstowych". I chociaż można polemizować z precyzyjnością powyższej definicji, to trudno nie zgodzić się co do jej istoty, według której metody i środki informatyki są niezwykle pomocne w selekcji i zestawianiu danych uzyskiwanych z dokumentów tekstowych, co znacznie ułatwia człowiekowi wyciąganie z tych zestawień wartościowych wniosków i spostrzeżeń.

Celem badań opisanych w niniejszym artykule jest wykorzystanie EAT do identyfikacji kluczowych obszarów rozważań teoretycznych dziedziny systemów informacyjnych zarządzania.

2. Hipoteza badawcza

Dziedziną, której dotyczą badania opisane w niniejszym artykule są systemy informacyjne zarządzania (SIZ). Ich centralnym przedmiotem zainteresowań jest, jak podają Lyytinen i King [2004, s. 221]: „«rynek idei», na którym naukowcy (i praktycy) wymieniają swoje poglądy dotyczące projektowania i zarządzania informacją oraz związanych z nią technik w zorganizowanym, ludzkim przedsiębiorstwie".

SIZ jako obszar interdyscyplinarny wyrasta z wielu pokrewnych dziedzin nauki i praktyki. Należy do nich zaliczyć przede wszystkim: ekonomię, socjologię, psychologię, nauki o zarządzaniu, informatykę i badania operacyjne [Laudon, Laudon 2002, s. 14].



Rys. 1. Systemy informacyjne zarządzania jako dziedzina interdyscyplinarna
Źródło: opracowanie własne na podstawie [Laudon, Laudon 2002, s. 14].

Naukowcy i praktycy w różny sposób akcentują poszczególne składowe. Możemy mówić zatem o podejściu tzw. twardym, lub inaczej technicznym, w którym dominują informatyka, i badania operacyjne, oraz tzw. miękkim, lub inaczej behawioralnym – zdominowanym przez socjologię i psychologię. Czynnikiem spajającym podejście twarde jest ogólna teoria systemów [Bertalanffy 1968], natomiast ogniwem łączącym podejście miękkie – miękka metoda systemowa (*soft system methodology*) [Checkland 1981; Checkland, Holwell 1998]. Relacje między poszczególnymi podejściami ilustruje rys. 1.

Chociaż historia SIZ liczy sobie już ponad trzydzieści lat, od pewnego czasu zauważa się głosy polemizujące z jej naukową legitymizacją. Jednym z pierwszych głosów owej polemiki był opublikowany ponad dziesięć lat temu artykuł Benbasata i Webera [1996]. Autorzy ci próbowali wykazać, że główną przeszkodą w uzyskaniu naukowej legitymizacji dziedziny SIZ jest jej interdyscyplinarność. Nie krytykowali oni jednak samej interdyscyplinarności SIZ. Wskazywali bowiem na fakt, że SIZ są dziedziną różnorodną zarówno co do rozważanych problemów, jak i podstaw teoretycznych oraz aspektów metodologicznych. Jednak głównym postulatem Benbasata i Webera był argument, że czerpanie inspiracji teoretycznych z dziedzin pokrewnych pozbawia SIZ własnego rdzenia teoretycznego. Dyskusję na temat znaczenia istnienia odrębnego rdzenia teoretycznego jako atrybutu naukowości dziedziny podjęli Lyytinen i King [2004]. Autorzy ci, stojąc na stanowisku, że żaden jeden spójny, wypracowany na własnym gruncie rdzeń teoretyczny nie legitymizuje dziedziny naukowej, osadzili swe rozważania m.in. w kontekście historii nauki. Wykazali, że czynnikami legitymizującymi naukowość danej dziedziny są: (1) istotność rozważanych problemów, (2) jakość rezultatów uzyskiwanych z prowadzonych badań oraz (3) plastyczność dziedziny, będąca odzwierciedleniem zdolności do podejmowania problemów badawczych oraz dawania na nie odpowiedzi w zmieniających się okolicznościach, zwłaszcza w wymiarze czasowym.

Niniejszy artykuł ma na celu określenie, jakie koncepcje i modele teoretyczne dziedzin pokrewnych, tj. ekonomii, psychologii, socjologii, nauk o zarządzaniu, informatyki i badań operacyjnych oraz w jakim stopniu stanowią rdzeń teoretyczny SIZ. Tak postawiony cel pozwala na sformułowanie hipotezy badawczej. Brzmi ona następująco: *Interdyscyplinarność SIZ jest wspierana przez multidyscyplinarność rozważanych teorii*. Jeśli dziedzina SIZ ma charakter interdyscyplinarny, w skład koncepcji teoretycznych rozważanych na jej gruncie powinny wchodzić koncepcje teoretyczne dziedzin pokrewnych zachowujące równomierne proporcje.

3. Materiał empiryczny

W celu weryfikacji postawionej hipotezy badawczej o interdyscyplinarności SIZ postanowiono wykorzystać teksty artykułów naukowo-badawczych opublikowanych w renomowanych czasopismach akademickich z dziedziny SIZ. Do wstępnej selekcji włączono dziewięć czasopism z dziedziny SIZ należących do Listy Filadelfijskiej, posiadających najwyższą punktację w wykazie czasopism opublikowanym przez Ministra Nauki i Informatyzacji z 7 października 2005 r.[1] oraz rankingu opublikowanego przez Thomson Scientific[2] wskaźnika *Impact Factor*. Rezultaty wstępnej selekcji zaprezentowano w tab. 1.

[1] http://www.nauka.gov.pl/mein/index.jsp?place=Lead08&news_cat_id=470&news_id=2959&layout = 2&page =text.

[2] http://scientific.thomson.com/.

Tabeli 1. Ranking czasopism z dziedziny SIZ z Listy Filadelfijskiej

Lp.	Tytuł	KBN	IF	Rok	Od
1	„MIS Quarterly"	24	4,731	2006	1977
2	„ACM Transactions on Information Systems"	24	4,097	2004	1983
3	„Information Systems Research"	24	2,054	2005	1990
4	„Information Systems"	24	1,887	2006	1976
5	„Information Systems Journal"	15	1,543	2006	1991
6	„Journal of Management Information Systems"	24	1,406	2005	1984
7	„Journal of Information Technology"	20	1,239	2006	1986
8	„Journal of Strategic Information Systems"	15	0,971	2005	1992
9	„European Journal of Information Systems"	20	0,862	2006	1992

Źródło: opracowanie własne.

Ze względu na rozmiary niniejszego opracowania postanowiono ograniczyć się do jednego, najbardziej renomowanego periodyku – „MIS Quarterly". Dokonując takiego właśnie wyboru, kierowano się następującymi względami: (1) „MIS Quarterly" jest uważany za niezaprzeczalnego lidera rankingów czasopism z dziedziny SIZ, (2) może poszczycić się już trzydziestoletnią historią oraz (3) jego charakter zawarty w misji kwartalnika[3]: „*Celem MIS Quarterly jest polepszenie przekazu wiedzy związanej z rozwojem usług wspartych IT, zarządzaniem zasobami technicznymi oraz ekonomią i zastosowaniami IT wraz z ich organizacyjnymi i menedżerskimi implikacjami*" w pełni oddaje interdyscyplinarność SIZ. Wydaje się zatem, że próba składająca się ze wszystkich artykułów badawczych opublikowanych w ciągu 30 lat wydawania „MIS Quarterly" będzie wystarczająco reprezentatywna do weryfikacji postawionej hipotezy badawczej.

Ponieważ artykuły naukowe są tekstami o dość rygorystycznej strukturze, stanowiącej swego rodzaju ontologię (strukturę znaczeniową), w prowadzonych badaniach postanowiono wykorzystać ich zwięzłą reprezentację w postaci słów kluczowych oraz streszczenia. Podejście takie znajduje uzasadnienie szczególnie w kontekście przeprowadzonych badań, których istotą jest poszukiwanie koncepcji teoretycznej stanowiącej podstawę prowadzonych rozważań. Świadome nawiązanie przez autora określonego artykułu do danego modelu teoretycznego powinno znaleźć odzwierciedlenie w umieszczeniu stosownego odniesienia w sekcji słów kluczowych lub przynajmniej w streszczeniu.

Przyjęte założenia doprowadziły do określenia próby badawczej składającej się z artykułów badawczych posiadających sekcję „słowa kluczowe" za okres 30 lat, tj. 1977-2006. Jako źródła danych wykorzystano w niej sekcje „słowa kluczowe" (*keywords*) oraz „streszczenie" (*abstract*). W ten sposób wyodrębniono 686 artykułów. W przeprowadzonych badaniach posłużono się pakietem statystycznym *Statistica 7.1*, w szczególności modułem *Text & Document Mining*, arkuszem kalkulacyjnym *MS Excel* oraz kilkoma autorskimi skryptami napisanymi w językach *Perl* i *bash*.

[3] http://www/misq.org/.

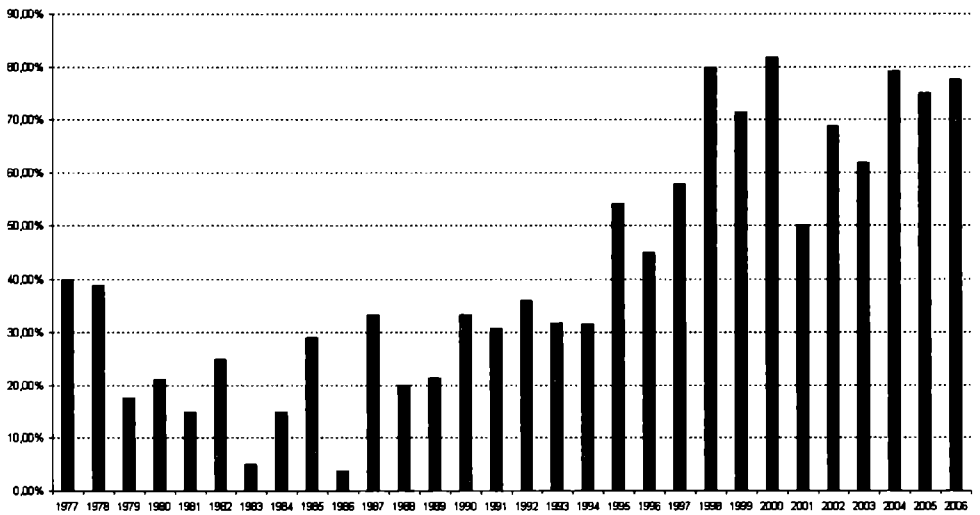
4. Rezultaty badań

W wyniku przeprowadzonych badań otrzymano następujące charakterystyki dotyczące użytych słów kluczowych. Ich łączna liczba w przeszukiwanych dokumentach wyniosła 3500, przy czym łączna liczba różnych użyć słów kluczowych to 2226. Oznacza to, że autorzy wykazywali się stosunkową dowolnością w definiowaniu słów kluczowych, ponieważ jedynie 36,4% użyć określonego słowa kluczowego stanowi jego ponowne użycie.

Przyjęto, że artykuł o charakterze teoretycznym to taki, który zawiera w sekcji *słowa kluczowe* i/lub *streszczenie* jeden z terminów: *theory*, *theoretic*, *theoretical*, *model*. Jeśli chodzi o samą sekcję *słowa kluczowe*, to użyć takich terminów było 129 (3,69%), przy czym różnych użyć terminów o charakterze teoretycznych w sekcji *słowa kluczowe* było 91 (4,09%).

Aby określić artykuły o charakterze teoretycznym, posłużono się następującą procedurą badawczą, stanowiącą klasyczne postępowanie badawcze w obszarze EAT: (1) dokonano redukcji do rdzenia języka angielskiego, (2) dokonano usunięcia nieistotnych słów przy użyciu stop-listy języka angielskiego, (3) zidentyfikowano istotne frazy – określono synonim słów *theory* oraz *theoretic*, *theoretical* i *model*; (4) w wyniku przeprowadzonej analizy otrzymano binarną postać macierzy częstości.

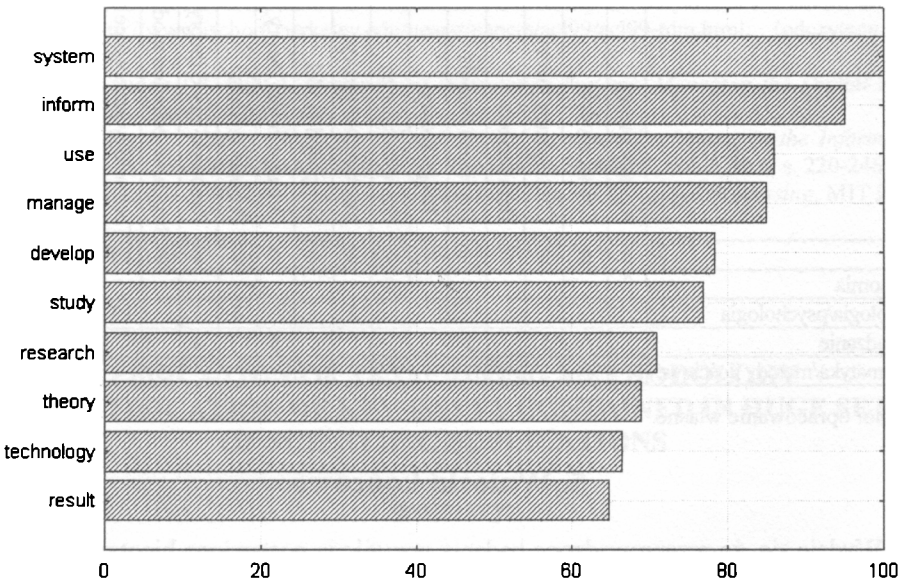
Wyniki przeprowadzonej analizy są zaprezentowane na rys. 2 i 3. Pierwszy z nich ilustruje procentowy udział artykułów o charakterze teoretycznym w ciągu 30 lat.



Rys. 2. Procentowy udział artykułów o charakterze teoretycznym
Źródło: opracowanie własne.

Rysunek 3 prezentuje istotność wyrażoną w procentach (100% najważniejszy – 0% nieważny) używanych terminów w analizowanym zbiorze danych. Wyniki

uzyskano dzięki zastosowaniu procedury SVD (*singular value decomposition*) binarnej macierzy częstości.



Rys. 3. Dziesięć najbardziej istotnych z 932 terminów
Źródło: opracowanie własne.

Analizując rys. 2 i 3, można wywnioskować, że począwszy od drugiej połowy lat 90., znacznie większą uwagę przywiązywano do teoretycznego kontekstu rozważań. Świadczy to niewątpliwie o procesie naukowego dojrzewania SIZ. Aspekt teoretyczny rozważań ma dużą rangę istotności (68,94%), zajmując 8 miejsce wśród 10 najważniejszych terminów.

W następnym kroku badawczym dokonano analizy istotności teorii szczegółowych. Procedura badawcza była podobna do poprzedniej, z wyjątkiem niedokonywania redukcji do rdzenia oraz braku stop-listy. W zamian zastosowano listę zawierającą jedynie słowa kluczowe o charakterze teoretycznym. Założono, że aby można było uznać określoną teorię lub model teoretyczny za istotny, musiał on wystąpić przynajmniej w 3 artykułach. W wyniku zastosowania procedury SVD określono istotność poszczególnych koncepcji teoretycznych. Z przeprowadzonej analizy wynika, że najbardziej popularną koncepcją badawczą jest model przyswojenia technologii (*technology acceptance model*) – 100% oraz teoria rozumnego działania (*theory of resoned action*) – 67,94%. Do najrzadziej stosowanych koncepcji teoretycznych należy teoria instytucjonalna (*institutional theory*) – 38,17% oraz modele decyzyjne (*decision models*) – 34,66%. W tab. 2 zestawiono liczbę artykułów prezentujących określoną koncepcję teoretyczną oraz przypisano je do określonych dziedzin rdzenia teoretycznego SIZ.

Tabela 2. Szczegółowe koncepcje teoretyczne

	Agency theory	Casual models	Contingency theory	Data models	Decisions models	Design theory	Economic theory	Institutional theory	Integrated model	Organizational theory	Process models	Resource-based theory	Structural equation modeling	Technology acceptance model	Theory of planned action	Theory of reasoned action	Transaction cost theory	
	3	**3**	**5**	**3**	**3**	**4**	**6**	**3**	**4**	**5**	**3**	**4**	**5**	**13**	**3**	**6**	**3**	
Ekonomia	3						6										3	**12**
Socjologia/psychologia								3	4		3			13	3	6		**32**
Zarządzanie			5		3					5		4						**17**
Informatyka/metody ilościowe		3		3		4							5					**15**

Źródło: opracowanie własne.

5. Wnioski końcowe

Wydaje się, że przeprowadzone badania weryfikują postawioną hipotezę, należy jednak zauważyć, że wśród omawianych koncepcji teoretycznych dominują teorie behawioralne, a szczególnie teoria rozumnego działania (*theory of reasoned action*) oraz jej rozwinięcie w postaci teorii planowego działania (*theory of planned behavior*) i jej szczególne zastosowanie w SIZ, tj. model przyswojenia technologii (*technology acceptance model*). Informatyka oraz metody ilościowe stanowią relatywnie mniejszy wkład teoretyczny.

Wraz z rozwojem dziedziny SIZ coraz większą uwagę przykłada się do jej podstaw teoretycznych. Stanowi to niewątpliwie dowód na coraz większą dojrzałość naukową dziedziny.

Metody eksploracyjnej analizy tekstu okazały się przydatne w rozważanym kontekście badawczym. Szczególnie użyteczna okazała się metoda SVD.

Literatura

Ackoff R.L. (1967), *Management Misinformation Systems*, „Management Science", December, vol. 14, nr 4, s. 147-156.

Benbasat I., Weber R. (1996), *Research Commentary: Rethinking „Diversity" in Information Systems Research*, „Information Systems Research", December, vol. 7, nr 4, s. 389-399.

Bertalanffy von L. (1968), *General System theory: Foundations, Development, Applications*, George Braziller, New York.

Checkland P. (1981), *Systems Thinking, Systems Practice*, Wiley, Chichester.

Checkland P., Holwell S. (1998), *Information, Systems and Information Systems*, Wiley, Chichester.
Hearst M. (2003), *What is Text Mining?,* October, http://www.sims.berkeley.edu/hearst/text-mining.html, (odczytano dn. 20.03.2007).
Hearst M. (1999), *Untangling Text Data Mining.* „Proceedings of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics", University of Maryland, June 20-26, (invited paper), http://www.ischool.berkeley.edu/hearst/papers/acl99/acl99-tdm.html, (odczytano dn. 20.03.2007).
Laudon K.C., Laudon J.P. (2002), *Management Information Systems. Managing the Digital Firm*, Prentice-Hall, Upper Saddle River,.
Lyytinen K., King J.L. (2004), *Nothing At The Center?: Academic Legitimacy in the Information Systems Field*, „Journal of the Association for Information Systems", vol. 5, nr 6, s. 220-246.
Manning C., Schütze H. (1999), *Foundations of Statistical Natural Language Processing*, MIT Press, Cambridge.
Witten I.H., Frank E. (2000), *Data Mining*, Morgan-Kaufmann, New York.

APPLICATION OF TEXT MINING METHODOLOGY TO IDENTIFICATION OF KEY ISSUES CONTAINED IN HUGE SETS OF SCIENTIFIC PUBLICATIONS

Summary

The development of data storage and retrieval tools, observed over the last decade, resulted in the increase of data availability – the text documents in particular. This increased availability enables the detailed analysis of huge sets of documents. The paper presents the use of text mining methods to identify the key theoretical issues in management information systems field. As a research sample abstracts and keywords will be used from all research papers published over the 30-year period (1997-2006) in top ranking research paper of management information systems field – *MIS Quarterly*.