PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
Nr 47 2009
TAKSONOMIA 16
Klasyfikacja i analiza danych – teoria i zastosowania

Mariusz Grabowski
Uniwersytet Ekonomiczny w Krakowie

WYKORZYSTANIE METOD EKSPLORACYJNEJ ANALIZY TEKSTU DO IDENTYFIKACJI ZMIENNOŚCI KONCEPCJI ZAWARTYCH W DUŻYCH ZBIORACH PUBLIKACJI NAUKOWYCH

1. Wstęp

Dynamiczny rozwój Internetu spowodował znaczne zwiększenie dostępności zbiorów danych, szczególnie dokumentów tekstowych. To niewątpliwie pozytywne zjawisko stwarza również wiele problemów w rozmaitych dziedzinach życia. Na przykład w przypadku działalności naukowej rzetelnie prowadzone badania w wielu przypadkach wymagają od badacza przeanalizowania niespotykanej dotąd liczby tekstów, co częstokroć przekracza jego możliwości percepcyjne.

W rozwiązaniu wyżej opisanego problemu pomocne okazują się metody i środki informatyki. Należy jednak zaznaczyć, że dokumenty tekstowe stanowią spore wyzwanie przy próbie ich automatycznego przetwarzania. Do czynników utrudniających to zadanie należy zaliczyć: brak spójnych reguł interpretacyjnych, niejednoznaczność terminologiczną (synonimy i homonimy) czy potencjalną wewnętrzną sprzeczność dokumentów tekstowych, wyrażającą się w kolokwialnym stwierdzeniu, że „papier zniesie wszystko". Jednak obecnie wspomniana zwiększona dostępność dokumentów stanowi nie tyle hamulec, ile motywację do opracowywania metod pozwalających na automatyzację przynajmniej pewnej części procesu interpretacji dokumentów tekstowych.

Dziedziną, której przedmiotem zainteresowania jest automatyzacja pozyskiwania informacji z dokumentów tekstowych, jest eksploracyjna analiza tekstu (EAT) [Hearst 1999] lub – zapożyczając brzmienie terminu z języka angielskiego – *text mining*. Przedmiot zainteresowań EAT można zdefiniować jako [Hearst 2003]: „odkrywanie przez komputer nowych, wcześniej nieznanych informacji, dzięki automatycznemu wydobywaniu informacji z różnych źródeł tekstowych". I chociaż można polemizować z jakością powyższej definicji, trudno nie zgodzić się co do jej istoty, według której metody i środki informatyki są pomocne w selekcji i ze-

stawianiu danych uzyskiwanych z dokumentów tekstowych, co znacznie ułatwia człowiekowi wyciąganie z tych zestawień wartościowych wniosków i spostrzeżeń.

2. Hipoteza badawcza

Niniejszy artykuł stanowi kontynuację badań opisanych w pracy [Grabowski 2008], a dotyczących wykorzystania EAT do weryfikacji hipotezy o naukowej legitymizacji dziedziny systemów informacyjnych zarządzania (SIZ).

Od połowy lat 90. w literaturze dotyczącej SIZ zauważa się głosy kwestionujące naukową legitymizację dziedziny bądź jej broniące. Do najbardziej znanych artykułów tego nurtu należy zaliczyć prace [Benbasat, Weber 1996] (po stronie krytyków) oraz [Lyytinen, King 2004] (po stronie obrońców).

Benbasat i Weber [1996] wskazywali, że główną przeszkodą w uzyskaniu naukowej legitymizacji dziedziny SIZ jest jej interdyscyplinarność oraz brak tzw. rdzenia teoretycznego, przy czym nie zdefiniowano, co oznacza to pojęcie. W pracy [Grabowski 2008] podjęto próbę określenia rdzenia teoretycznego dziedziny SIZ jako teorii zapożyczonych i rozwijanych na jej gruncie. Do rdzenia teoretycznego zaliczono następujące teorie szczegółowe: teorię 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*), teorię kosztów transakcyjnych (*tarnsaction cost theory*), teorię agencyjną (*agency theory*) oraz inne teorie perspektywy ekonomii, teorie kontyngencji (*contingency theory*) oraz teorię zasobową (*resource based theory*). Wśród artykułów o wyraźnie zaznaczonych szczegółowych inspiracjach teoretycznych dominuje perspektywa behawioralna.

W odpowiedzi na tzw. dyskurs niepokoju [Benbasat, Weber 1996] Lyytinen i King [2004] 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 zweryfikowanie trzeciego z postulatów Lyytinena i Kinga. Plastyczność dziedziny SIZ będzie tutaj rozumiana jako zmienność koncepcji rozważanych w dziedzinie SIZ w całej jej historii, tj. w ciągu ostatnich 30 lat.

3. Materiał empiryczny

W celu weryfikacji postawionej w poprzednim punkcie hipotezy badawczej zgromadzono zbiór streszczeń[1] artykułów naukowych opublikowanych w reno-

[1] W niektórych przypadkach wykorzystanie streszczenia artykułu zamiast jego pełnej wersji przynosi lepsze rezultaty. Do takich przypadków można zaliczyć np. identyfikację kluczowych zagadnień zawartych w dużych zbiorach publikacji naukowych. W innych zadaniach EAT, jak np. klasyfikacja i grupowanie dokumentów, trudno jest jednoznacznie określić, jaka forma jest lepsza, choć pewną przewagę wykazuje pełny tekst [Zheng i in. 2005; Cohen i in. 2005].

mowanych czasopismach akademickich dotyczących dziedziny SIZ. Za kryterium określające rangę danego czasopisma przyjęto punktację zawartą w dokumentach Ministra Nauki i Szkolnictwa Wyższego [MNiSW 2007a; MNiSW 2007b].

W stosunku do badań opisanych w artykule [Grabowski 2008], opartych na analizie streszczeń, które ukazały się w latach 1977-2006 na łamach „MIS Quarterly", zbiór empiryczny został rozszerzony o cztery tytuły. Zabieg ten miał na celu zwiększenie reprezentatywności próby badawczej w wymiarze ilościowym przez dołączenie artykułów z innego periodyku amerykańskiego oraz przekrojowym przez dołączenie periodyków europejskich. W rezultacie przeprowadzonych modyfikacji zbiór empiryczny zawiera streszczenia 2351 artykułów opublikowanych w latach 1997-2006 w pięciu periodykach naukowych dziedziny SIZ: dwóch amerykańskich i trzech europejskich[2].

Nazwy czasopism, wartości punktacji MNiSW, wartości wskaźnika *Impact Factor* za rok 2007, lata rozpoczęcia publikacji, liczbę artykułów oraz miejsce (kraj) publikacji zestawiono w tab. 1.

Tabela 1. Czasopisma naukowe w zbiorze empirycznym

Lp.	Tytuł	MNiSW	IF	Od	Liczba	Miejsce
1	„MIS Quarterly"	30	4.731	1977	691	USA
2	„Information Systems Research"	24	2.537	1990	378	USA
3	„Journal of Information Technology"	24	1.239	1986	645	Wlk. Brytania
4	„Information Systems Journal"	20	1.543	1991	191	Wlk. Brytania
5	„European Journal of Information Systems"	20	0.862	1992	446	Wlk. Brytania

Źródło: opracowanie własne na podstawie [MNiSZ 2007b] oraz JCR 2006 [Internet 1].

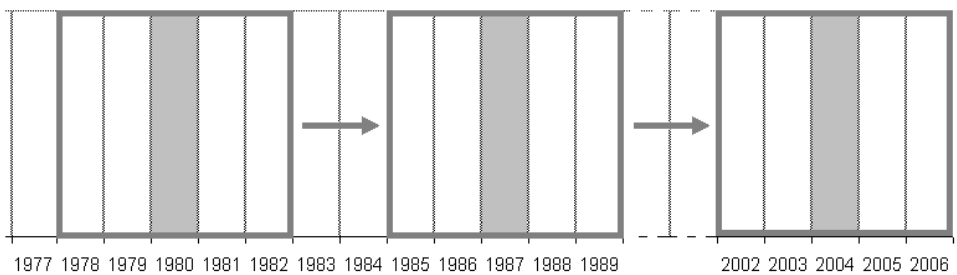
Chociaż trudno jest zweryfikować empirycznie reprezentatywność opisanego wyżej zbioru dokumentów, to jednak wydaje się, że zbiór ten w sposób zadowalający prezentuje kompleksowość problematyki w 30-letniej historii SIZ i nadaje się do weryfikacji hipotezy o zmienności rozważanych koncepcji dziedzinowych, zawiera bowiem dwa najbardziej prestiżowe periodyki amerykańskie oraz trzy periodyki europejskie o najwyższych wskaźnikach *Impact Factor* oraz najwyższej punktacji MNiSW [Benbasat, Zmud 2003; Katerattankul, Han, 2003]. Również procentowy udział artykułów amerykańskich (46%) oraz europejskich (54%) jest zbliżony.

4. Procedura badawcza

W celu uwzględnienia wymiaru czasowego analizowanego zbioru dokumentów podzielono go na 30 podzbiorów, w których znalazły się artykuły opublikowane w określonych latach (1977-2006). Jednak ze względu na to, że okres jednego roku wydaje się zbyt krótki, aby przyjąć go za podstawową jednostkę analizy, postanowiono przyjąć okres pięcioletni. Aby w jak największym stopniu odzwierciedlić

[2] Ze względu na rangę oraz język publikacji są to wyłącznie periodyki brytyjskie.

dynamikę koncepcji, zrezygnowano z całkowicie arbitralnego podziału badanego okresu na pięcioletnie podokresy. W zamian zastosowano znaną w prognozowaniu szeregów czasowych koncepcję ruchomego okna. Szerokość okna obejmuje pięć lat. Okno takie reprezentuje rok leżący w jego środku (po jego lewej stronie znajdują się dwa lata poprzedzające, a po prawej dwa lata następujące).



Rys. 1. Metoda przesuwnego okna

Źródło: opracowanie własne.

Okno przesuwane jest o jeden rok od momentu, gdy jego lewa krawędź umiejscowiona jest na początku okresu[3], tj. w roku 1978, aż do momentu, gdy jego prawa krawędź znajdzie się na roku 2006. W ten sposób analizie będzie poddane 25 okresów o środkach od roku 1980 do roku 2004.

Dla każdego okresu przeprowadzona jest analiza ważności koncepcji przy użyciu metody LSA (*Latent Semantic Analysis*) [Deerwester i in. 1990] realizowanej z wykorzystaniem algorytmu SVD (*Singular Values Decopmposition*) [*Metody statystycznej…* 2004]. W badaniach skorzystano z implementacji zawartej w pakiecie *Statistica 8.0*. Poniżej opisano szczegółową procedurę badawczą.

1. Przyjęto, że koncepcje są odzwierciedlone przez terminy, a te z kolei składają się ze słów. Zgodnie z podstawowym założeniem EAT o braku informacji *a priori* o analizowanych tekstach nie definiowano fraz ani synonimów;

2. Dokonano redukcji do rdzenia języka angielskiego oraz zastosowano mechanizm listy wyłączeń języka angielskiego. Do macierzy częstości włączono określony termin jedynie w przypadku, jeśli wystąpił w co najmniej 3% plików. Zabieg ten ma na celu wykluczenie z analizy terminów przypadkowych;

3. Analizę przeprowadzono dla wszystkich dostępnych postaci macierzy częstości zaimplementowanych w module *text mining* pakietu *Statistica 8.0*, tj. oryginalnej ($\mathbf{X}$), binarnej ($\mathbf{X}^{\mathbf{bin}}$), logarytmicznej ($\mathbf{X}^{\mathbf{log}}$) i ważonej logarytmicznej ($\mathbf{X}^{\mathbf{idflog}}$). W dalszej części niniejszego opracowania poszczególne postacie macierzy częstości będą nazywane wariantami metody SVD;

[3] Ze względu na spójność analizy oraz niewielką dostępność danych w początkowych okresach analizy zdecydowano się dołączyć dane z roku 1997 do okna o środku umiejscowionym w roku 1980.

4. Dokonano uporządkowania (malejącego) ważności terminów według wskaźnika ważności terminu opisanego wzorem:

$$\hat{w}_i^{rel} = \frac{\hat{w}_i}{\hat{w}_1},$$

gdzie: $\hat{w}_i$ to ważność *i*-tego terminu otrzymanego za pomocą metody SVD, $\hat{w}_1$ to ważność najważniejszego terminu otrzymanego za pomocą metody SVD. Tak znormalizowany wskaźnik przyjmuje wartości w przedziale $< 0; 1 >$.

5. Dokonano modyfikacji niektórych słów do formy rzeczownikowej dla zwiększenia czytelności określonego terminu [Abramowicz 2008].

W trakcie badań przyprowadzono 100 analiz (25 analiz dla każdej postaci macierzy częstości).

5. Rezultaty badań

W tabelach 2 i 3 zamieszczono rankingi najważniejszych terminów w poszczególnych okresach wraz z wartościami miernika $\hat{w}_i^{rel}$. Wyniki uzyskane przy użyciu

Tabela 2. Najważniejsze koncepcje SI w ujęciu dynamicznym (na podstawie macierzy X)

1980		1981		1982		1983		1984	
system	1,00	system	1,00	system	1,00	system	1,00	system	1,00
information	0,67	information	0,65	information	0,70	information	0,69	information	0,78
management	0,59	management	0,64	management	0,64	management	0,62	management	0,60
mis	0,51	mis	0,50	development	0,52	development	0,51	development	0,51
1985		1986		1987		1988		1989	
system	1,00	system	1,00	system	1,00	system	1,00	system	1,00
information	0,85	information	0,85	information	0,84	information	0,86	information	0,89
management	0,63	management	0,57	management	0,58	management	0,59	management	0,59
development	0,54	decision	0,50	decision	0,56	decision	0,51	decision	0,52
1990		1991		1992		1993		1994	
system	1,00	system	1,00	system	1,00	system	1,00	system	1,00
information	0,86	information	0,89	information	0,90	information	0,95	information	0,94
management	0,60	use	0,64	user	0,70	model	0,75	model	0,81
use	0,54	management	0,63	model	0,65	user	0,72	user	0,72
1995		1996		1997		1998		1999	
system	1,00	system	1,00	information	1,00	system	1,00	system	1,00
information	0,98	information	1,00	system	0,98	information	0,98	information	0,97
model	0,82	model	0,89	model	0,87	model	0,83	model	0,81
user	0,71	user	0,65	research	0,64	research	0,69	research	0,75
2000		2001		2002		2003		2004	
system	1,00	information	1,00	information	1,00	information	1,00	information	1,00
information	0,96	system	0,97	system	0,97	system	0,98	system	0,98
model	0,82	research	0,78	research	0,89	research	0,88	research	0,87
research	0,80	model	0,75	model	0,74	knowledge	0,86	knowledge	0,81

Źródło: opracowanie własne.

metod $\mathbf{X}^{\mathbf{bin}}$ oraz $\mathbf{X}^{\mathbf{log}}$ są bardzo zbliżone do tych, które zostały uzyskane przy użyciu metody **X**. We wszystkich analizowanych okresach metody te za najważniejsze uznały terminy *information* oraz *system*, w niektórych przypadkach zamienione jedynie miejscami. Dlatego biorąc pod uwagę ograniczenia niniejszego opracowania, zamieszczono jedynie rezultaty uzyskane przy użyciu dwóch wariantów metody SVD: **X** oraz $\mathbf{X}^{\mathbf{idflog}}$. Ograniczono się przy tym jedynie do zaprezentowania jedynie czterech najważniejszych pojęć.

Tabela 3. Najważniejsze koncepcje SI w ujęciu dynamicznym (na podstawie macierzy X^{idflog})

1980		1981		1982		1983		1984	
audit	1,00	dss	1,00	dss	1,00	dss	1,00	dss	1,00
dss	0,98	word	0,95	job	0,73	language	0,78	graphical	0,87
office	0,89	network	0,85	network	0,69	job	0,75	language	0,72
decision	0,89	office	0,82	office	0,68	strategic	0,68	decision	0,71
1985		1986		1987		1988		1989	
dss	1,00	decision	1,00	decision	1,00	decision	1,00	decision	1,00
decision	0,86	dss	0,97	group	0,81	group	0,92	group	0,94
data	0,75	group	0,87	dss	0,79	user	0,82	user	0,85
job	0,75	support	0,81	support	0,77	support	0,81	software	0,82
1990		1991		1992		1993		1994	
group	1,00	group	1,00	dss	1,00	expert	1,00	group	1,00
decision	0,89	expert	0,97	group	0,99	group	0,97	expert	0,87
software	0,82	user	0,93	user	0,94	user	0,90	user	0,83
user	0,81	knowledge	0,92	expert	0,93	software	0,85	job	0,81
1995		1996		1997		1998		1999	
group	1,00	outsourcing	1,00	outsourcing	1,00	outsour-	1,00	outsour-	1,00
learn	0,85	group	0,94	group	0,97	software	0,99	knowledge	0,89
user	0,85	software	0,82	decision	0,84	group	0,98	software	0,88
model	0,85	decision	0,81	investment	0,81	decision	0,90	group	0,88
2000		2001		2002		2003		2004	
outsourcing	1,00	erp	1,00	erp	1,00	knowled	1,00	trust	1,00
software	0,89	team	0,97	outsourcing	0,98	outsour-	0,96	erp	0,98
knowledge	0,86	outsourcing	0,95	trust	0,94	trust	0,93	knowledge	0,98
group	0,81	knowledge	0,95	knowledge	0,88	erp	0,90	software	0,95

Źródło: opracowanie własne.

Analizując dane zawarte w tab. 2 i 3, można zauważyć, że wyniki uzyskane dzięki metodzie $\mathbf{X}^{\mathbf{idflog}}$ zdają się potwierdzać hipotezę o plastyczności dziedziny SIZ. Plastyczność dziedzinowa wyraża się w zmienności rozważanych koncepcji. Początkowo, w latach 80., dominowała problematyka wspomagania decyzji (*decision*, *dss*), przy czym w drugiej połowie lat 80. zyskała ona wymiar zespołowy (*group*). Początek kolejnej dekady w jeszcze większym stopniu akcentował problematykę zespołowego podejmowania decyzji (*group*), rozwijana była przy tym problematyka systemów doradczych (*expert*). Druga połowa lat 90. wraz z postępującą globalizacją

zwróciła uwagę na zjawisko outsourcingu IT (*outsourcing*). W pierwszej połowie pierwszej dekady XXI wieku dziedzina SIZ przeniosła główny punkt zainteresowania na problematykę systemów klasy ERP (*erp*), zarządzania wiedzą (*knowledge*) oraz zagadnień związanych z zaufaniem (*trust*), które nabrało szczególnego znaczenia wraz z dynamicznym rozwojem biznesu elektronicznego. Chociaż metoda $\mathbf{X}^{\mathbf{idflog}}$ wydaje się w największym stopniu odzwierciedlać zmienność koncepcji w rozpatrywanym okresie, to należy zaznaczyć, że również pozostałe warianty ($\mathbf{X}$, $\mathbf{X}^{\mathbf{bin}}$, $\mathbf{X}^{\mathbf{log}}$) odzwierciedlają pewne trendy. Nie tyle określają one rozważaną problematykę, ile definiują kontekst rozważań. Można np. zauważyć, że w początkowym okresie rozwoju dziedziny SI (lata 80.) ważne miejsce zajmowały kwestie związane z zarządzaniem (*management*). Nieco później akcen towano aspekty behawioralne, kładąc nacisk na kwestie związane z użyciem (*use*). W latach 90. zwracano uwagę na problematykę związaną z modelowaniem (*model*), a przełom wieków i początkowe lata obecnego stulecia przyniosły duże zainteresowanie problematyką badawczą (*research*). Wobec tego można wyciągnąć wniosek, że dziedzina SIZ podlega ewolucji. Przenosi akcent z kwestii praktycznych na teoretyczno-badawcze, o czym świadczy awans w rankingu pojęcia *research*, a spadek pojęcia *management*. Potwierdza to niewątpliwie rozwój dziedziny jako dyscypliny akademickiej.

Zaprezentowana w niniejszym artykule metoda może być również pomocna w określaniu istotnych problemów badawczych. Dzięki niej można np. wzbogacić funkcjonalność oprogramowania bibliotecznego stanowiącego platformę dostępową do publikacji naukowych.

Literatura

Abramowicz W. (2008), *Filtrowanie informacji*, AE, Poznań.

Benbasat I., Weber R. (1996), *Research commentary: rethinking "diversity" in information systems research*, „Information Systems Research", December vol. 7, no 4, s. 389-399.

Benbasat I., Zmud R.W. (2003), *The identity crisis within the is discipline: defining and communicating the discipline's core properties*, „MIS Quarterly" vol. 27, no 2, s. 183-194.

Cohen A.M., Yang J., Hersh W.R. (2005), *A comparison of techniques for classification and ad hoc retrieval of biomedical documents*, „NIST Special Publication: SP 500-266", The Fourteenth Text Retrieval Conference, http://trec.nist.gov/pubs/trec14/papers/ohsu-geo.pdf.

Deerwester S., Dumlas S.T., Furnas G.W., Landauer T.K., Harshman R. (1990), *Indexing by latent semantic analysis*, „Journal of The American Society for Information Science" vol. 41, no 6, s. 391-407.

Grabowski M. (2008), *Wykorzystanie metod eksploracyjnej analizy tekstu do identyfikacji kluczowych zagadnień zawartych w dużych zbiorach publikacji naukowych*, [w:] K. Jajuga, M. Walesiak (red.), Taksonomia 15, *Klasyfikacja i analiza danych – teoria i zastosowania*, AE, Wrocław.

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.

Hearst M. (2003), *What is text mining?,* October, http://www.sims.berkeley.edu/ hearst/text-mining.html.

Katerattankul P., Han B. (2003), *Are European journal under-rated? An answer based on citation analysis*, „European Journal of Information Systems", vol. 12, s. 60-71.

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, no 6, s. 220-246.

Metody statystycznej analizy wielowymiarowej w badaniach marketingowych (2004), red. E. Gatnar, M. Walesiak, AE, Wrocław.

MNiSW (2007a), *List Ministra w sprawie listy czasopism punktowanych, Zn. DBB/4901/2007*, Ministerstwo Nauki i Szkolnictwa Wyższego, Warszawa, http://www.nauka.gov.pl/mn/_gAllery/32/66/32663/20071119_list_ministra.pdf.

MNiSW (2007b), *Wykaz wybranych czasopism wraz z liczbą punktów za umieszczoną w nich publikację naukową*, Ministerstwo Nauki i Szkolnictwa Wyższego, Warszawa, http://www.nauka.gov.pl/mn/_gAllery/32/66/32664/20071119_Wykaz_czasopism.pdf.

Zheng Z.H., Brady S., Garg A., Shatkay H. (2005), *Applying probabilistic thematic clustering for Classification in the TREC 2005 Genomics Track*, „NIST Special Publication: SP 500-266", The Fourteenth Text Retrieval Conference, http://trec.nist.gov/pubs/trec14/papers/queensu.geo.pdf.

Źródła internetowe [1] http://scientific.thomson.com/products/jcr /.

THE APPLICATION OF TEXT MINING METHODOLOGY IN THE IDENTIFICATION OF PLASTICITY OF CONCEPTS CONTAINED IN LARGE SETS OF SCIENTIFIC PUBLICATIONS

Summary

The paper presents the application of text mining methodology in the identification of the concept plasticity in the field of information systems over the period of 30 years. A procedure that links the moving window technique with LSA algorithm (*latent semantic analysis*) [Deerwester et al. 1990] is proposed. Abstracts from all electronically available articles (2348), published in five renowned scientific journals of information systems domain: *MIS Quarterly*, *Information Systems Research*, *Journal of Information Technology*, *Information Systems Journal* and *European Journal of Information Systems* are used as a database for the research.