AUTOMATYKA • 2011 • Tom 15 • Zeszyt 2

Edward Nawarecki*, Stanisława Kluska-Nawarecka**,
Joanna Dziaduś-Rudnicka**, Dorota Wilk-Kołodziejczyk***

Aspekty technologiczne i systemowe pozyskiwania informacji z otwartych źródeł

1. Wstęp

Od szeregu lat w Instytucie Odlewnictwa, przy udziale Katedry Informatyki oraz w Katedrze Informatyki Stosowanej i Modelowania AGH, prowadzone są prace dotyczące komputerowego udostępniania wiedzy technologicznej pozyskiwanej z rozproszonych źródeł [1, 2, 3, 4, 14].

Biorąc pod uwagę coraz większe znaczenie jakie odgrywają w świecie współczesnym niezmierzone zasoby informacyjne sieci Internet [5, 6, 7], oczywiste jest, że informacje te powinny stanowić jedną z ważnych składowych wiedzy dotyczącej w szczególności nowych technologii, surowców i materiałów, komponentów procesu technologicznego, potencjalnych zamówień, itp.

Jak wiadomo, w powszechnym użyciu znajduje się szereg wyszukiwarek internetowych, które w przypadku „codziennych potrzeb informacyjnych" okazują całkowicie wystarczające. Równocześnie jednak w przypadku potrzeb o charakterze bardziej zaawansowanym, gdy chodzi o informacje z pewnego (nie zawsze dobrze zdefiniowanego) obszaru problemowego, lub monitorowanie informacji posiadających charakter dynamiczny, możliwości oferowane przez dostępne wyszukiwarki stają się nie wystarczające do efektywnego rozwiązania danego problemu.

W artykule przedstawiono rezultaty pewnego etapu prac eksperymentalnych dotyczących pozyskiwania informacji o technologiach odlewniczych ze stron WWW, których wyniki posłużyły do stworzenia koncepcji oraz podjęcia działań realizacyjnych komputerowego systemu prowadzącego proces eksploracji i klasyfikacji danego typu informacji w sposób zautomatyzowany. Podano wybrane rezultaty eksperymentów, przedstawiono schemat funkcjonalny systemu oraz naszkicowano niektóre procedury obliczeniowe.

* AGH Akademia Górniczo-Hutnicza, Wydział Elektrotechniki, Automatyki, Informatyki i Elektroniki, Katedra Informatyki
** Instytut Odlewnictwa, Kraków
*** Krakowska Akademia im. Andrzeja Frycza Modrzewskiego

2. Pozyskiwanie informacji o technologiach odlewniczych ze stron WWW

Przeprowadzone eksperymenty miały na celu uzyskanie rozeznania w zakresie efektywności wykorzystania ogólnie dostępnych wyszukiwarek internetowych (m. i. Google, Onet, Yahoo……) do pozyskania informacji, a w konsekwencji wiedzy o charakterze technologicznym. Biorąc pod uwagę potrzeby prowadzonych aktualnie prac badawczych, jak również fakt, że realizacja tego typu eksperymentu wymaga głębszego rozeznania obszaru problemowego, którego ma on dotyczyć, skoncentrowano się na poszukiwaniu informacji dotyczących technologii odlewniczych (także materiałów dla odlewnictwa).

W wyniku przeprowadzonych eksperymentów pozyskano obszerny materiał poglądowy, z którego starano się wydobyć pewne cechy charakteryzujące proces poszukiwania. Zgodnie z przewidywaniami okazało się, że wyniki zastosowania wyszukiwarek w dużym stopniu mają charakter losowy. Równocześnie jednak pozwalają na uzyskanie wskaźników ilościowych, które w pewnym stopniu umożliwiają formułowanie wniosków o charakterze porównawczym.

Ponieważ w wielu przypadkach liczba stron wytypowanych przez wyszukiwarkę może być bardzo duża, analizie poddana została tylko ich część (próbka losowa). W wyniku analizy przeprowadzonej przez eksperta, następuje wskazanie stron „trafionych", czyli takich, których treść odpowiada oczekiwaniom użytkownika.

Głównym celem eksperymentu było zbadanie wpływu zastosowanych zestawów słów kluczowych na efekty poszukiwania, oceniane na podstawie liczby zaindeksowanych stron WWW oraz trafności dokonanego wyboru, mierzonego współczynnikiem „sukcesu" danego wyrażenia kluczowego:

$$\eta_i = \frac{w_{it}}{w_{i0}},$$

gdzie:

w_{it} – liczba stron trafionych, dla i-go słowa kluczowego,
w_{i0} – liczba stron ocenianych, dla i-go słowa kluczowego.

Charakterystyczny fragment prowadzonych analiz przedstawiono w formie tabeli 1.

Określony w powyższy sposób współczynnik sukcesu posłużył do stworzenia rankingu słów kluczowych (widoczny w tab. 1), który prowadzono osobno dla wyszukiwania „zaawansowanego" oraz „ prostego".

Z uzyskanych danych liczbowych wynika, że liczba „trafionych" stron, wskazanych przez eksperta prawie liniowo zależy od liczności próbki. Tak więc przyjęcie liczności próbki rzędu kilkudziesięciu stron (tutaj 20, 40), można uznać za wystarczające do uzyskania reprezentatywnej oceny.

Podobne rezultaty otrzymano dla zestawów słów kluczowych charakteryzujących inne obszary technologii odlewniczych.

Tabela 1
Wybrane rezultaty wyszukiwania stron WWW

Lp.	Słowa kluczowe	Rodzaj wyszukiwania	Liczba znalezionych stron	Liczba analizowanych stron	Współczynnik sukcesu (η)	Skuteczność wyszukiwania
1	piankowy filtr ceramiczny „odlewnictwo”	zaawansowane	108	40	0,125	dość dobra
2	piankowy filtr ceramiczny „staliw”	zaawansowane	86	20	0,05	słaba
3	staliw „filtry piankowe”	zaawansowane	163	40	0,025	słaba
4	staliw „piankowy filtr ceramiczny”	zaawansowane	15	15	0,00	słaba
5	staliw„filtry ceramiczne”	zaawansowane	175	40	0,00	słaba
6	piankowe filtry ceramiczny dla odlewnictwa	proste	2890	40	0,375	dobra
7	piankowy filtr ceramiczny	proste	3620	40	0,275	dobra
8	piankowe filtry ceramiczne dla staliw	proste	1340	20	0,20	dobra

Przykładowe zestawy słów kluczowych charakteryzujących inne obszary technologii odlewniczych:

„rafinatory do stopów Al” – $\eta = 0{,}25$,
„materiały dla odlewnictwa” – $\eta = 0{,}00$.

Jak widać z przedstawionych danych, pełna selekcja uzyskanego z wyszukiwarki materiału, czyli wskazanie wszystkich stron zawierających interesujące użytkownika informacje, wymaga analizy treści tekstu zawartego na wszystkich zaindeksowanych stronach.

Wiadome jest, że analiza semantyczna dokumentów tekstowych (szczególnie dla języka fleksyjnego jakim jest j. polski) jest zadaniem bardzo trudnym, które nie znalazło dotychczas w pełni zadawalającego rozwiązania. Dlatego racjonalne wydaje się zastosowanie metod heurystycznych, niewymagających skomplikowanych narzędzi programistycznych, ani zbyt dużego nakładu obliczeń. Jedną ze stosunkowo prostych i często stosowanych tej klasy metod jest określenie wektorowej reprezentacji analizowanego tekstu [5, 8].

W badanym dokumencie określana jest liczba tych samych wyrazów występujących w tekście, a w konsekwencji tworzony jest wektor:

$$W = [w_1, w_2, ... w_n]$$

gdzie w_i – jest liczbą wystąpień i-tego wyrazu na stronie.

W rozważanym przypadku, zliczanie wszystkich słów wydaje się zbędne, szczególnie, że dokumenty poddane zostają wstępnej selekcji przez wyszukiwarkę.

Proponuje się, aby zawartość wektora W ograniczyć do składowych określających liczbę wystąpień dominujących zestawów wyrazów.

Kolejna seria eksperymentów dotyczyła efektywności zastosowania tego podejścia w obszarze technologii odlewniczych.

Przykładem otrzymanych rezultatów, może być ocena wektorowa stron zaindeksowanych przy zastosowaniu słowa kluczowego: **filtry ceramiczne piankowe**.

Dla strony ocenionej przez eksperta jako „trafiona" otrzymano wektor:

$$W = [filtr - 12,\ stop - 10,\ \ odlew - 5,\ \ staliw - 3,\ \ żeliw - 2]$$

dla strony „nietrafionej" wyznaczono odpowiednio:

$$W' = [filtr - 3,\ stop - 3,\ \ odlew - 0,\ \ staliw - 0,\ \ żeliw - 0]$$

Jak widać, wektory W oraz W' różnią się w sposób istotny, co w danym przypadku wskazuje na skuteczność wektorowej oceny zawartości analizowanych stron.

Podobnie obiecujące wyniki uzyskano dla stron zaindeksowanych przy użyciu innych zestawów słów kluczowych.

Przedstawione rezultaty stanowią oczywiście tylko wstępne rozeznanie w zakresie charakterystyki procesu pozyskiwania stron Internetowych, zawierających informacje o technologiach odlewniczych. Wydaje się jednak, że mogą być przyjęte za podstawę do sformułowania planu dalszych badań oraz stworzenia koncepcji systemu informatycznego wspierającego te badania.

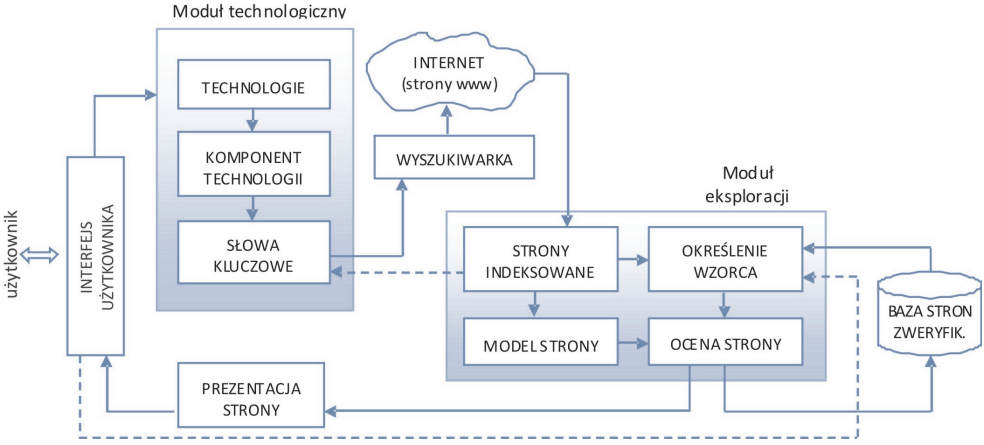
3. Koncepcje rozwiązań systemowych

Jako punkt wyjścia do tworzenia koncepcji systemu informatycznego, wspierającego pozyskiwanie informacji ze stron WWW, przyjęto następujące stwierdzenia wynikające z opisanych powyżej eksperymentów:

– powszechnie stosowane wyszukiwarki internetowe, działające w oparciu o podawane zestawy słów kluczowych, dokonują indeksacji stron WWW z dużym nadmiarem – tylko niewielka część tych stron (10–20%) zawiera informacje istotne dla użytkownika;

– wskazywanie „trafionych" stron przez eksperta jest żmudne i czasochłonne, czyli w konsekwencji potrzebne jest stworzenie mechanizmów automatycznego wyboru stron udostępnianych użytkownikowi;
– wstępne wyniki eksperymentów wskazują, że zastosowanie wektorowej oceny zawartości stron powinno okazać się skuteczne, pod warunkiem opracowania narzędzi programistycznych umożliwiających tworzenie wzorców modeli stron dla poszczególnych kategorii (obszarów) informacji technologicznych.

Schemat funkcjonalny konstruowanego systemu wspomagającego pozyskiwanie informacji technologicznych ze stron WWW został przedstawiony na rysunku 1.



Rys. 1. Schemat funkcjonalny systemu eksploracji wiedzy technologicznej

Moduł technologiczny systemu przeznaczony jest do sformułowania zadania, realizowanego z udziałem użytkownika, który wskazuje rodzaj technologii stanowiącej przedmiot jego zainteresowania. Na tej podstawie użytkownikowi zostaje udostępniony wykaz komponentów danego procesu (materiały, urządzenia, wyposażenie, ...), co umożliwia doprecyzowanie wskazań dotyczących obszaru poszukiwań (doboru stron WWW).

Kolejny blok zawiera katalog słów kluczowych, odpowiadających przewidywanym obszarom eksploracji oraz uszeregowanym zgodnie z przyjętym rankingiem, określającym kolejność ich stosowania.

Wyjściem modułu technologicznego jest wybrany zestaw słów kluczowych, który wprowadzony jest do wyszukiwarki internetowej. Zarówno wyszukiwarka, jak i sieć Internet działają poza projektowanym systemem, zaś wynikiem tego działania jest zbiór zaindeksowanych stron, udostępnianych na wejściu modułu eksploracji.

Moduł eksploracji stanowi kluczową część projektowanego systemu, gdzie dokonywane jest określenie modelu wektorowego poszczególnych stron, konstrukcja wzorca, a następnie ewaluacja stron.

Model wektorowy strony definiowany jest jako zestaw słów (wyrazów) z odpowiednimi wagami, określającymi krotność ich występowania na danej stronie.

Do modelu zaliczane są tylko te wyrazy, których waga jest niemniejsza od zadanego progu. Ponieważ liczba słów, które zaliczone zostaną do modelu jest *a priori* nieznana, dlatego stosuje się określenie progu w procentach najwyższej wagi w danym modelu, czyli:

$$W = [w_1\ slowo\,1,\ w_2\ slowo\,2, ..., w_n\ slowo\,n], \qquad \text{dla}\ \ w_1 \geq w_2 ..... \geq w_n$$

przy czym:

$$w_n \geq \delta\%\ \max_{i=1}^{N} w_i$$

gdzie:

N – liczba wszystkich słów danej strony podlegających zliczaniu,

n – liczba słów włączonych do modelu ($n \leq N$),

w_i – waga (krotność wystąpienia) i-go słowa ($i = 1,..... n$).

Ponieważ liczba słów na stronach, a w konsekwencji liczba wag w_i może być bardzo różna, dlatego zachodzi potrzeba normalizacji wektora W, czyli obliczenie:

$$\overline{w_i} = \frac{w_i}{\sum_{i=1}^{n} w_i}$$

Wówczas model wektorowy przybiera postać:

$$\overline{W} = [\overline{w}_1\ slowo\,1, ..., \overline{w}_n\ slowo\,n], \quad \sum_{i=1}^{n} \overline{w}_i = 1$$

Kolejny blok modułu eksploracji dokonuje definiowania wzorca W^*, z którym porównywany będzie model danej strony.

Istnieje wiele metod tworzenia wzorców służących do klasyfikacji obiektów określonych przez odpowiednio zdefiniowany zbiór atrybutów. Do tej klasy metod należą: uczenie maszynowe, algorytmy ewolucyjne, sieci neuronowe.

Tutaj, uznano jako zasadne zastosowanie jednego z algorytmów eksploracji danych (*data mining*) [9, 10, 11, 12, 13], gdzie poszukuje się wzorców częstych.

Stosując to podejście, zbiór analizowanych stron traktowany jest jako zbiór transakcji, których atrybutami są poszczególne wyrazy występujące na danej stronie.

Rozważając zbiór możliwych wzorców:

$$W^* = \left\{ W_1^*, W_2^* ....., W_j^* \right\}$$

poszukuje się takiego, dla którego krotność występowania będzie większa od założonej wartości minimalnego wsparcia:

$$\sup(W_i^*) > \sup_{\min}^{k}$$

gdzie: $\sup_{\min}^{k}$ – minimalne wsparcie, dla którego wymagana jest co najmniej *k*-krotne wystąpienie wzorca W_i^*.

Wyznaczone zostają wzorce (W_i^*) dla pewnej sekwencji wartości minimalnego wsparcia, czyli $k = \alpha, \alpha+1,....., \alpha+m$ zawierające coraz to mniej wyrazów, przy czym $\alpha+m$ odpowiada sytuacji, gdy nie znaleziono żadnego wyrazu występującego w analizowanych stronach w krotności $\alpha+m$ razy.

Porównując wektory uzyskane dla kolejnych wartości *k*, można wyznaczyć wzorzec W^* zawierający wyrazy w liczbach określających wagi w_i poszczególnych słów

$$W^* = \left[(\alpha + m - 1)\ wyraz\ 1, (\alpha + m - 2)\ wyraz\ 2, ... \alpha\ wyraz\ m \right]$$

Po wprowadzeniu normalizacji składowych wektora W_i^*, określony zostaje znormalizowany wektor wzorcowy:

$$\overline{W}^* = [\ \overline{w}_1\ wyraz\ 1, ..., \overline{w}_m\ wyraz\ m], \qquad \sum_{i=1}^{m} \overline{w}_i^* = 1$$

Z wzorcem z tym porównywane są wektory $\overline{W}$ wyznaczone dla kolejnych stron przy użyciu wybranej metryki np. odległości euklidesowej. Strona zaklasyfikowana zostaje jako „trafiona" przy spełnieniu warunku:

$$\left\| \overline{W}^* - \overline{W}_i \right\| \le \varepsilon$$

gdzie ε – dopuszczalna wartość błędu.

Oprócz dopisanych powyżej modułów, na schemacie z rysunku 1 zaznaczono bloki o charakterze pomocniczym, do których należą;

- **użytkownik**, który umożliwia prowadzenie dialogu z modułem technologicznym oraz udostępnienie wyników eksploracji;
- **prezentacja stron** polega na sprowadzeniu ich do formy dogodnej do interpretacji przez użytkownika (np. przez wyeksponowanie najważniejszych elementów strony);
- **baza stron zweryfikowanych** umożliwia archiwizację stron, które mogą być w przyszłości wykorzystane bezpośrednio – to jest z pominięciem całej procedury eksploracji.

Informacyjne sprzężenia zwrotne, zaznaczone na schemacie linią przerywaną, wskazują potencjalną możliwość adaptacji odnośnych procedur, poprzez wykorzystanie doświadczeń wynikających z poprzednio zrealizowanych procesów eksploracji.

Należy zwrócić uwagę, że opisany schemat z rysunku 1 ma charakter funkcjonalny, a więc nie musi odpowiadać „fizycznej" architekturze systemu, która wynika z przyjętych rozwiązań implementacyjnych.

4. Uwagi końcowe

W pracy rozważany jest problem pozyskiwania wiedzy technologicznej przy wykorzystaniu otwartych źródeł informacji, ze szczególnym uwzględnieniem stron WWW.

Przedstawiono wybrane rezultaty wstępnych eksperymentów dotyczących obszaru technologii odlewniczych. Wnioski z tych eksperymentów, w połączeniu z oglądem dostępnych narzędzi i rozwiązań informatycznych, stały się podstawą przedstawionej koncepcji systemu, który łączy aspekty technologiczne z informatycznymi, obejmującymi w szczególności metody eksploracji danych i tworzenia wzorców. Aktualnie, niektóre z bloków tego systemu znajdują się w fazie testowania, zaś inne wymagają jeszcze dopracowania założeń projektowych.

Reasumując, przedstawiony w pracy materiał, z jednej strony stanowi relację z pewnego etapu prowadzonych prac, z drugiej zaś określa dalsze zamierzenia badawcze i realizacyjne.

Praca finansowana przez Ministerstwo Nauki i Szkolnictwa Wyższego w ramach projektu, decyzja nr 820/N-Czechy/2010/0.

Literatura

[1] Górny Z., Kluska-Nawarecka S., Wilk-Kołodziejczyk D., Regulski K., *Diagnosis of casting defects using uncertain and incompleteknowledge.* Archives of Metalurgy and Materials, nr 3, 2010, 819–826.

[2] Kluska-Nawarecka S., Wilk-Kołodziejczyk D., Dobrowolski G., Nawarecki E., *Strukturalization of knowledge about casting defects diagnosis based on rough sets theory*. Computer Methods in Materials Science, vol. 9, No 2, 2009.

[3] Kluska-Nawarecka S., Dobrowolski G., Marcjan R., Nawarecki E., *Od pasywnych do aktywnych źródeł danych i wiedzy: zdecentralizowany system informacyjno-decyzyjny dla wspomagania technologii odlewniczej*. Wyd. AGH, Kraków 2002.

[4] Kluska-Nawarecka S., Regulski K., *Zarzadzanie wiedzą w systemach wspomagania technologii materiałowe.* Zastosowanie Systemów, nr 36, 73–86, Kraków 2007.

[5] Brzeziński J., Morzy M., Morzy T., *Algorytmy optymalizacji zapytań eksploracyjnych z wykorzystaniem materializowanej perspektywy eksploracyjnej.* http://www.cs.put.poznan.pl/mmorzy/papers/rb006-02.pdf .

[6] Wilk-Kołodziejczyk D., *Pozyskiwanie wiedzy w sieciach komputerowych z rozproszonych źródeł informacji.* winntbg.bg.agh.edu.pl/skrypty2/0095/285-295.pdf.

[7] Macioł A., Stawowy A., Wrona R.A., *Web based foundry knowledge base*. Archives of Foundry engineering Polish Academy of Science, vol. 9, Issue 1, 2009, 5–8.
[8] Kowalik P., *Wspomaganie tworzenia analizatorów stron wyników z internetowych systemów wyszukujących.* http://www.cs.put.poznan.pl/mmorzy/papers/rb006-02.pdf.
[9] Mirończuk M., *Eksploracja danych w kontekście procesu Knowledge Discovery In Databases (KDD) i metodologii Cross-Industry Standard Process For Data Mining*. Metody Informatyki Stosowanej, 2009, http://pan.wi.zut.edu.pl/pliki/MIS-2009-2.
[10] Bodon E.F., *Frie-based apriori implementation for mining frequent itensequences*. Procedings of ACM SIGKDD International Workshop on Open Source Data Mining Chicago, USA, 2005, 56–65.
[11] Han J., Pei J., Yin Y., Mao R., *Mining frequent patterns without candidate generation: A frequent – patterns free approach.* Data Min. Knowl. Discov., 8(1), 2004, 53–87.
[12] Seifert J.W., *Data mining: An overview*. CRS Report for Congress, 2004.
[13] i2Ltd,i2 Pattern Tracer, http://www.12.co.uk/, 2009.
[14] Dobrowolski G., Marcjan R., Nawarecki E., Kluska-Nawarecka S., Dziaduś J., Wójcik T., *Development of INFOCASTT: Information system for Fouundry Industry*. Task Quartely 7, no 2, 2003.