PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 1121 — 2006
Pozyskiwanie wiedzy i zarządzanie wiedzą

Jakub Swacha, Roman Budzowski, Artur Kulpa

Uniwersytet Szczeciński

GENERATOR ZESTAWIEŃ DANYCH LICZBOWYCH

1. Zestawienia danych

Zestawienie danych przedstawia dane dowolnego typu w postaci ukierunkowanej na zwięzłość, główne punkty zainteresowania, stosunkowo umiarkowaną estetykę i ogólne dążenie do ujęcia jak największej ilości istotnej informacji w jak najmniejszej ilości tekstu. Jako takie stanowią one sposób prezentacji informacji idealnie pasujący do ponowoczesnego wieku informacji (por. [Furmanek 2005]), kiedy to dane mnożą się z coraz większą szybkością, podczas gdy maleje czas przeznaczony na zapoznanie się z nimi i zrozumienie ich. W pewnych przypadkach mogą służyć jako podsumowania, tudzież streszczenia dłuższego dokumentu.

Powyższy opis, oparty na definicji zaczerpniętej z [*Reference*... 2005], dobitnie wyjaśnia, dlaczego zestawienie danych jest tak interesującym rozwiązaniem w potopie informacji dzisiejszego świata. Gdy chodzi o szybkie znalezienie odpowiedzi, użytkownika zadowoli raczej jasno podany suchy fakt, a nie pięknie ułożone zdanie.

Wymienione wyżej zalety stanowią mocny powód do tworzenia zestawień danych. Nie ma jednak potrzeby, by zestawienia danych wykonywano ręcznie. Sposób, w jaki zestawienia danych mogą być generowane automatycznie z tekstu w języku naturalnym, zostanie opisany w dalszej części niniejszego artykułu.

Koncepcja generowanych automatycznie zestawień danych

Pożądana możliwość szybkiego zapoznania się z ogólną treścią dokumentu tekstowego bez potrzeby czytania go w całości była głównym powodem licznych projektów systemów automatycznego podsumowywania dokumentów. W ciągu niecałych 50 lat, które upłynęły od wydania wczesnych prac P.H. Luhna [1959], zaprezentowano szereg metod ekstrakcyjnych [Edmundson 1964; Hovy, Chin-Yew 1997; Kupiec i in. 1995] i generacyjnych [DeJong 1982; Radev, McKeown 1998; Rau, Jacobs, Zernick 1989].

Zadaniem systemów automatycznego podsumowywania dokumentów jest dostarczenie użytkownikowi informacji, o czym ogólnie traktuje streszczany tekst. Dlatego celem autorów systemów automatycznego podsumowywania dokumentów jest, by wynik ich działania zbliżony był do streszczeń pisanych przez wykwalifikowanych ludzi [Jing, McKeown 2000].

Proponowane w niniejszym artykule ujęcie ma zupełnie inny charakter. Zadaniem wygenerowanych automatycznie zestawień danych jest dostarczenie użytkownikowi wyciągu danych zawartych w tekście.

W przekonaniu autorów, w niejednej sytuacji taki wyciąg może dla odbiorcy mówić więcej niż ogólne streszczenie. Ponadto, jako że te dwa rodzaje podsumowań wzajemnie się dopełniają, ich jednoczesne udostępnienie czytelnikowi wydaje się być nader praktycznym pomysłem.

Mimo że metodologicznie automatyczne generowanie zestawień danych stanowi odrębną koncepcję, z technicznego punktu widzenia jest ono bliskie ekstrakcyjnym metodom automatycznego podsumowywania dokumentów. Jednak od wygenerowanego automatycznie zestawienia danych nie ma potrzeby wymagać ani tego, by pokrywało ono najbardziej istotne tematy dokumentu, ani tego, by w jakikolwiek sposób udawało, że zostało przygotowane przez człowieka. Oba te fakty rozluźniają wymagania co do generowanej zawartości, w ten zaś sposób czynią cały pomysł łatwiejszym do zaprojektowania i realizacji. W przekonaniu autorów, jeżeli wygenerowane automatycznie zestawienie danych zawiera informacje bezwartościowe dla odbiorcy, mogą one zostać zignorowane przez niego bez wielkiego nakładu wysiłku, jako że całkowita liczba pozycji w zestawieniu jest niewielka. I choć wydaje się nieuniknione, że dla pewnych dokumentów wygenerowane automatycznie zestawienia danych będą bezużyteczne, to jest to do przyjęcia, gdyż dla wielu innych mogą się one okazać bardzo wartościowe.

Dane ujęte w zestawieniach

Zestawienia mogą ujmować różne rodzaje danych. Typy danych, które najszybciej przychodzą na myśl to nazwy (imiona, nazwiska, nazwy firm, towarów, itd.); wielkość, ilość i wartość; data i czas; miejsce (lokalizacja). Istnieją metody, rozwinięte przede wszystkim na potrzeby wyszukiwania i ekstrakcji informacji, które są w stanie odnaleźć w tekście dane dowolnego z wymienionych typów [Friburger, Maurel 2001; Jacobs, Rau 1990; Smith 2002; Smith, Crane 2001].

Pojawia się zatem pytanie, jakiego typu dane najbardziej zasługują na to, by pojawić się w generowanym automatycznie zestawieniu. Zdaniem autorów są to dane liczbowe (wielkość, ilość i wartość, a także data). Zdania zawierające liczby są informacyjne i precyzyjne. Tworzenie zestawienia na podstawie danych o charakterze liczbowym ma tę ważną zaletę, iż różnego rodzaju nazwy są zwykle używane jako słowa kluczowe w wyrażeniach przeszukiwania tekstu, podczas gdy liczby stanowią raczej odpowiedzi na takie zapytania. To stwierdzenie ma silne

oparcie w rzeczywistych danych: np. jedna z popularnych wyszukiwarek internetowych zgłosiła tylko jedną liczbę wśród 1000 najczęściej wyszukiwanych słów kluczowych [*Top 1,000...* 2005]. Poprzez ujęcie danych liczbowych w zestawieniu umożliwia się użytkownikowi łatwy dostęp do informacji, która mogłaby zostać przeoczona w zwykłym wyszukiwaniu.

Wykorzystanie zestawień danych

Tworzone w sposób automatyczny zestawienia danych mogą być wykorzystane na różne sposoby, zwłaszcza jako:

- uzupełnienie lub zamiennik streszczenia,
- narzędzia wspomagające wyszukiwanie,
- zestawienia danych jako takie.

W pierwszym z wymienionych zastosowań zestawienie istotnych danych stanowi doskonałe uzupełnienie streszczenia, ponieważ nie tylko pozwala czytającemu zorientować się w treści czytanego dokumentu, ale również daje mu przegląd danych, jakie ten dokument zawiera.

Zestawienia danych są bardzo pomocne w sytuacjach, kiedy kryteria wyszukiwania są trudne lub niemożliwe do określenia. Odpowiada to sytuacji, w której czytający wie, że dokument zawiera interesujące go dane, ale nie wie dokładnie, jakiego rodzaju są to dane.

Stworzone w automatyczny sposób zestawienia danych, jako takie, mogą być bardzo pomocne dla autorów lub redaktorów, którzy chcieliby umieścić w swoim dokumencie zestawienie danych. Ułatwiają również autorom, redaktorom czy recenzentom odkrycie błędów w zamieszczonych w dokumencie danych, poprzez ich wyeksponowanie.

2. Tworzenie zestawień danych liczbowych

Tworzenie zestawień danych liczbowych wymaga wyselekcjonowania istotnych danych z przetwarzanego dokumentu. Ponieważ nie da się jednoznacznie określić precyzyjnego zestawu cech określającego daną jako właściwą do ujęcia w zestawieniu, zdecydowano się na podejście heurystyczne.

Projekt heurystyki

W celu określenia reguł decydujących o istotności danych wybrano dwadzieścia artykułów z internetowego katalogu GoArticles [*Article Search...* 2005]. Wybrane artykuły zostały podzielone na dwa zbiory: uczący i testowy – każdy składający się z dziesięciu artykułów. Zbiór uczący został zanalizowany w celu określenia reguł pozwalających na automatyczne oszacowanie istotności przetwarzanych danych. Na podstawie wyników tej analizy zbudowano zbiór dwudziestu reguł, które okazały się wystarczające do prawidłowego przetworzenia uczącego zestawu artykułów.

Spośród dwudziestu odkrytych reguł wybrano siedem najbardziej użytecznych:

1) informacja jest traktowana jako istotna, jeżeli składa się z cyfr, separatora dziesiętnego i separatorów tysięcznych;

2) informacja jest istotna, jeżeli jest zapisana w jednym z powszechnych formatów zapisu daty;

3) informacja jest istotna, jeżeli jest słowną reprezentacją liczby;

4) informacja jest nieistotna, jeżeli jest mieszaniną cyfr i liter;

5) informacja jest nieistotna, jeżeli jest odwołaniem do literatury;

6) informacja jest nieistotna, jeżeli jest numerem pozycji listy;

7) informacja jest nieistotna, jeżeli jest zakwalifikowana jako numer lub odwołanie do elementu struktury dokumentu (np. rozdziału, strony, równania, tabeli, ilustracji itp.).

Tabela 1. Przykłady istotnych i nieistotnych liczb

Dane istotne	Dane nieistotne
1 100.00	Chapter 3
October 20th 2005	123a1332
$3 m	[Kowal 2000]
fifteen years	1. The books
15122	page 15
3 km	fig. 1

Źródło: opracowanie własne.

Każde zdanie zawarte w dokumencie i zawierające co najmniej jedną istotną daną jest traktowane jako istotne. W tabeli 1 zaprezentowano przykłady istotnych i nieistotnych liczb.

Algorytm generowania zestawień danych liczbowych

Algorytm generowania zestawień danych liczbowych składa się z pięciu głównych etapów:

1) wczytanie analizowanego tekstu;

2) podział tekstu na listę zdań;

3) ocena istotności danych;

4) ekstrakcja istotnych zdań;

5) wygenerowanie sformatowanego zestawienia.

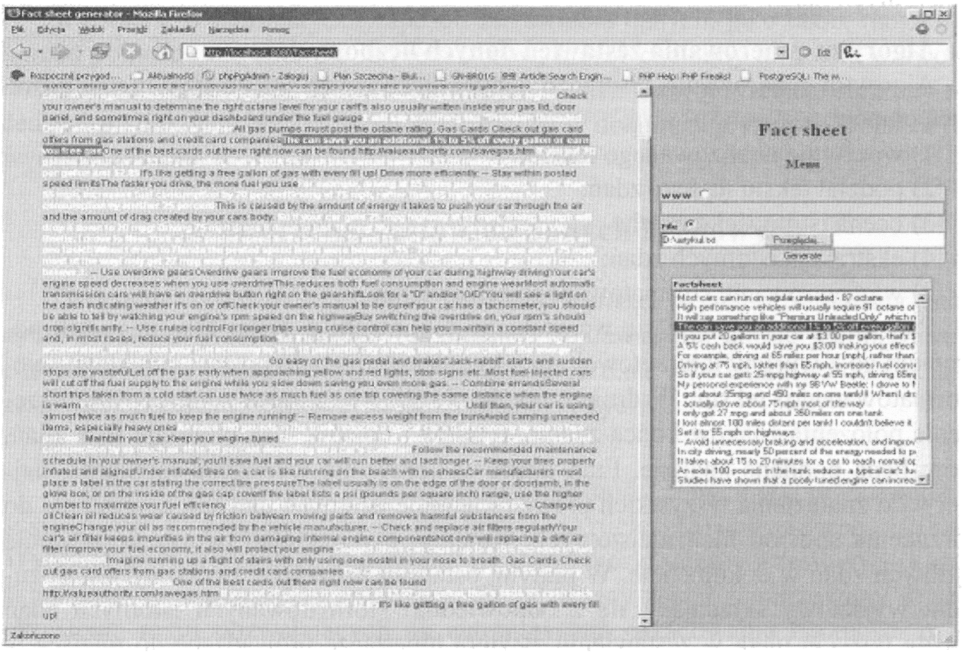
Pierwszy etap zajmuje się wczytaniem dokumentu z dysku lokalnego lub pobraniem go z witryny internetowej i przekształceniem go do postaci zwykłego (niesformatowanego) tekstu. W drugim etapie określa się granice zdań poprzez określenie położenia znaków końca zdania, tytułów i opisów. Tak powstaje lista zdań, które będą poddane analizie. Etap trzeci wykorzystuje wyrażenia regularne, najpierw do znalezienia wszystkich prawdopodobnie istotnych danych, a następnie do odrzucenia spośród nich nieistotnych. Kolejny, czwarty etap tworzy zestawienie istotnych zdań w dokumencie. W piątym, ostatnim już etapie, tworzony jest dokument wyjściowy składający się z odpowiednio sformatowanych istotnych zdań (każda pozycja ujęta w zestawieniu zawiera hiperłącze prowadzące do miejsca w oryginalnym tekście, w którym ją znaleziono) oraz oryginalnego tekstu z wyróżnionymi zdaniami zawierającymi istotne dane.

3. Generowanie zestawień danych w praktyce

Omówiony algorytm został zaimplementowany w postaci programu Automatyczny Generator Zestawień. Program ten posłużył do wygenerowania zestawień danych dla dziesięciu artykułów z zestawu testowego. Uzyskane w testach rezultaty zostały omówione poniżej.

Założenia dotyczące oprogramowania

Program Automatyczny Generator Zestawień został napisany w języku PHP [*PHP*... 2005] w architekturze klient-serwer. Interfejs użytkownika, zdefiniowany z wykorzystaniem języków HTML [Pfaffenberger, Karow 2001] i JavaScript [Goodman 2000], wykorzystuje do komunikacji z użytkownikiem przeglądarkę WWW (zob. rys. 1). Funkcjonalność i prostota interfejsu użytkownika pozwalają w łatwy sposób wczytać przeznaczony do przetworzenia tekst z dysku lokalnego lub pobrać go z zasobów internetowych. Wynik procesu analizy tekstu jest przedstawiany w dwóch ramkach. Pierwsza zawiera widok samego tekstu wraz z zaznaczonymi zdaniami zawierającym istotne dane, druga zaś właściwe zestawienie danych, wygenerowane na podstawie tekstu. Jeżeli użytkownik zaznaczy którąś z pozycji zestawienia, aplikacja automatycznie odnajdzie w ramce zawierającej analizowany tekst odpowiadający jej akapit, pozwalając użytkownikowi na zapoznanie się z szerszym kontekstem umieszczonych w zestawieniu danych.



Rys.1. Rezultat działania programu Automatyczny Generator Zestawień

Źródło: opracowanie własne.

Wyniki badań

Testowy zbiór tekstów wykorzystanych w eksperymencie zawierał dziesięć wybranych losowo artykułów: naukowych, gazetowych i analiz rynkowych. Za pomocą programu Automatyczny Generator Zestawień dla każdego z tych artykułów wygenerowano autòmatycznie zestawienie. Szczegółowe wyniki uzyskane w eksperymencie przeprowadzonym na omawianych tekstach zostały zamieszczone w tab. 2. Pierwsza kolumna zawiera identyfikator artykułu na stronie internetowej GoArticles [*Article Search...* 2005], z której wszystkie one zostały pobrane. Adres internetowy URL każdego z wymienionych artykułów można otrzymać z poniższego wzorca: http://www.goarticles.com/cgi-bin/showa.cgi?C=identyfikator, przez zastąpienie słowa „identyfikator" odpowiednim numerem artykułu.

Druga kolumna zawiera liczbę zdań zakwalifikowanych przez algorytm jako zawierające istotne dane i umieszczonych jako pozycje w zestawieniu. W trzeciej kolumnie znajduję się liczba zdań, które powinny były zostać zakwalifikowane jako istotne i dodane do zestawienia, lecz zostały pominięte przez algorytm. Czwarta kolumna zawiera liczbę zdań, które – choć nie są istotne – zostały włączone przez algorytm do zestawienia. W piątej, ostatniej kolumnie, znajduje się miara efektywności zaimplementowanego algorytmu, określona przez proporcję pomiędzy liczbą zdań istotnych ujętych w wygenerowanym zestawieniu a sumą liczby zdań w dokumencie zawierających istotne dane (zarówno ujętych, jak i pominiętych w wygenerowanym zestawieniu) i liczby zdań bez istotnych danych, a zakwalifikowanych do ujęcia w zestawieniu.

Wszystkie dane w zestawieniu zostały zweryfikowane przez człowieka. Na tej podstawie oceniania była prawidłowość zakwalifikowania istotnych danych oraz wychwytywane były dane, które zostały przez program pominięte.

Tabela 2. Efektywność działania programu Automatyczny Generator Zestawień

Identyfikator	Znalezione	Pominięte	Nieprawidłowe	Efektywność [%]
75306	8	0	0	100,00
76410	4	0	0	100,00
77323	5	1	0	83,33
77643	4	0	0	100,00
77910	4	1	0	80,00
79486	39	9	0	81,25
79620	3	0	0	100,00
79908	18	0	1	94,44
79976	7	1	0	87,50
80246	25	0	0	100,00
140763	2	0	0	100,00
140554	5	0	0	100,00
140720	6	1	0	85,71
Średnia	10	1	0,08	93,25
Odchylenie standardowe	–	–	–	8,23

Źródło: opracowanie własne.

4. Podsumowanie

Przedstawione w artykule zasady przetwarzania tekstu i oparty na nich algorytm pozwalają na automatyczne wygenerowanie zestawienia danych liczbowych na podstawie tekstu napisanego w języku naturalnym. Opisany algorytm zaimplementowano w postaci programu Automatyczny Generator Zestawień w wersji dostosowanej do tekstu w języku angielskim i do anglosaskiej notacji liczb i dat. Następnie przetestowano go na próbnym zestawie dokumentów, osiągając średnią efektywność wyszukiwania istotnych danych liczbowych na poziomie ponad 90%.

Generator zestawień danych liczbowych jest zarówno ideowo, jak i obliczeniowo mniej złożony od algorytmów tworzenia automatycznych podsumowań tekstu. Wygenerowane zestawienia mogą zostać użyte jako wyciąg treści dokumentu (uzupełniający streszczenie), jako narzędzie wspomagające wyszukiwanie danych, mogą być także przydatne w redagowaniu i recenzowaniu dokumentu.

Opisany algorytm generuje zestawienia danych typu liczbowego, które wybrano jako najbardziej nadające się do prezentacji w formie zestawienia. Nic nie stoi na przeszkodzie rozwinięciu algorytmu przez włączenie doń reguł pozwalających na generowanie zestawień danych innych typów.

Literatura

Article Search Engine Directory, http://www.goarticles.com/index.html, update October 15th (2005).

DeJong G.F., *An Overview of the FRUMP System. Strategies for Natural Language Processing*, Lawrence Erlbaum Associates, Hillsdale, NJ 1982, s. 149-176.

Edmundson H.P., *Problems in Automatic Extracting*, "Communications of the ACM" 1964 nr 7, s. 259-263.

Friburger N., Maurel D., *Finite-State Transducer Cascade to Extract Proper Names in Texts. Revised Papers from the 6th International Conference on Implementation and Application of Automata*, 2001, s. 115-124.

Furmanek M., *Kultura informacyjna w ponowoczesnym świecie*, [w:] *Edukacja informacyjna – technologie informacyjne w ponowoczesnym świecie*, red. K. Wenta, E. Perzycka, Wyd. CDiDN, Szczecin 2005, s. 37-42.

Goodman, D., *JavaScript. Księga eksperta*, Helion, Gliwice 2000.

Hovy E., Chin-Yew L., *Automated Text Summarization in SUMMARIST*, *Proc. ACL97/EACL97 Workshop on Intelligent Scalable Text Summarization*, Madrid, Spain 1997, s. 18-24.

Jacobs P.S., Rau L.F., *SCISOR: Extracting Information from On-Line News*, "Communications of the ACM" 1990 nr 33(11), s. 88-97.

Jing H., McKeown K., *Cut and Paste Based Text Summarization*, *Proc. of the 1st Meeting of the North American Chapter of ACL*, 2000, s. 178-185.

Kupiec J., Pedersen J., Chen F., *A Trainable Document Summarizer. Research and Development in Information Retrieval*, ACM 1995, s. 68-73.

Luhn P.H., *Automatic Creation of Literature Abstracts*, "IBM J. Res. Develop." 1959 nr 2, s. 159-165.

Pfaffenberger B., Karow B., *HTML 4. Biblia*, Helion, Gliwice 2001.
PHP: Hypertext Preprocessor, www.php.net, Update October 13th (2005).
Radev D., McKeown K., *Generating Natural Language Summaries from Multiple Online Sources*. "Computational Linguistics" 1998 nr 24(3), s. 469-500.
Rau L.F., Jacobs P.S., Zernick U., *Information Extraction and Text Summarization Using Linguistic Knowledge Acquisition*, "Information Processing and Management" 1989 nr 25(4), s. 419-428.
Reference.com Encyclopedia, http://www.reference.com/browse/wiki/Fact sheet, 2005.
Smith D.A., *Detecting Events With Date And Place Information In Unstructured Text. Proceedings of the 2nd ACM+IEEE Joint Conference on Digital Libraries*, Portland, OR 2002, s. 191-196.
Smith D.A., Crane G., *Disambiguating Geographic Names in a Historical Digital Library*, *Proc. of the 5th European Conference on Research and Advanced Technology for Digital Libraries*, 2001, s. 127-136.
Top 1,000 Searches, http://www.search.com/top?tag=se.fd.topnav.top, week ending October 2nd 2005.

GENERATOR OF NUMERIC DATA LISTS

Summary

The paper describes an algorithm for automatic generating numeric data lists based on a natural language text. The numeric data lists generated by the described algorithm consist of document sentences containing essentially relevant numbers. The sentence selecting algorithm is based on heuristics. Its usefulness has been verified by its software implementation and running tests on a sample text corpus. The automatically produced numeric data lists may be used as document digests (thus being useful for authors, editors, and readers), and also in applications like information retrieval and extraction as a search-supporting tool, especially when the user has not specified the correct search keyword.