Anna GŁADYSZ

PRZEGLĄD ZASTOSOWAŃ ANALIZY TEXT MININGOWEJ

W artykule omówiona została eksploracyjna analiza danych tekstowych ze szczególnym naciskiem na zastosowania analizy text miningowej. We współczesnym świecie istnieje wiele różnych branż biznesowych w których pracownicy stykają się z nadmiarem napływających informacji. Rozwój społeczeństwa informacyjnego oraz technologii informatycznych pociągnął za sobą w sposób naturalny powstanie zautomatyzowanych systemów wspomagających wyszukiwanie i porządkowanie informacji. Techniki text miningu znajdują coraz większe zastosowanie, zaś szeroki przegląd zastosowań wraz ze wskazaniem praktycznym możliwości zastosowania analizy text moningowej został dogłębnie omówiony w artykule.

WSTĘP

Eksploracyjna analiza tekstów (*text mining*) jest stosunkowo młodą multidyscyplinarną dziedziną. Wywodzi się z data mining, wyszukiwania informacji, ekstrakcji danych, kategoryzacji tekstu, modelowania probabilistycznego, wykorzystująca miedzy innymi metody statystyczne, czy maszynowe uczenie [1]. Pojęcie text miningu swoją popularność zdobywa od końca lat dziewięćdziesiątych XX wieku. Jego twórczynią była Marti A. Hearst, która zdefiniowała text mining jako „proces mający na celu odkrycie przez komputer nowych, poprzednio nieznanych informacji z zasobów tekstowych" [2]. Celem tego procesu jest odnalezienie istotnej, wcześniej nieznanej i wyczerpującej wiedzy z nieustrukturalizowanych danych tekstowych. Kluczowe znaczenie ma przy tym łączenie znalezionych informacji w całość dającą nowe fakty, czy hipotezy do dalszych badań. Biorąc pod uwagę różnorodność źródeł pozyskiwania danych tekstowych, a także łatwą ich „produkowalność" rola eksploracyjnej analizy tekstów będzie coraz bardziej istotna zarówno dla osoby prywatnej jak również dla efektywnego zarządzania organizacją.

Zakres zastosowań analizy text miningowej jest bardzo szeroki. Próbę określenia najważniejszych obszarów zastosowań można znaleźć w wielu pozycjach literaturowych [3]. Zaprezentowane w artykule zestawienie stanowi przegląd najważniejszych zastosowań text miningu.

1. EKSPLORACYJNA ANALIZA TEKSTU

1.1. Text mining

Text mining jest procesem analizowania naturalnie występującego tekstu w celu odkrywania i zapisania semantycznej informacji z ostatecznym celem jakim jest odkrycie wiedzy przez albo tekstowy albo wizualny dostęp do użycia w szerokim zasięgu ważnych aplikacji. Jest on uważany za podspecjalność szerszej domeny odkrywania wiedzy z danych (*Knowledge Discovery from Data, KDD*), który może zostać zdefiniowany jako proces obliczeniowy ekstrahowania przydatnej informacji z masywnych ilości cyfrowych danych przez mapowanie niskopoziomowe danych do pełniejszych, bardziej abstrakcyjnych form i przez wykrywanie istotnych wzorców pośrednio obecnych w danych [4]. Motywację dla rozpatrywania tego typu narzędzi stanowi ciągły wzrost technicznych możliwości gromadzenia i analizy danych, w których ukryte są potencjalnie cenne informacje dopełniające wiedzę ludzką. W wielu sytuacjach, stosowanie procesu KDD jest wręcz koniecznością, wynikającą chociażby z dramatycznie szybkiego przyrostu danych. Pełny proces KDD uwzględnia przechowywanie danych i dostęp do nich, oczyszczenie i przygotowanie danych, wykrywanie wzorców, przekształcanie i redukcję danych, eksplorację danych oraz zastosowanie określonych algorytmów w celu dostrzeżenia i ekstrahowania wzorców. Text mining wzbogaca radykalnie przydatność KDD wykorzystując przetwarzanie języków naturalnych. NLP dostarcza niezbędne metody dla text miningu, by automatycznie wydobywać wiedzę z tekstów.

1.2. Historia text miningu

H. P. Luhn w 1958 roku w nowatorskim artykule na temat automatycznego tworzenia abstraktów zanotował „przeanalizować władzę istotnych słów w tekście źródłowym" [5]. Lauren B. Doyle w 1961 również utrwalił istotę text miningu i metod pokrewnych, kiedy powiedział, że „naturalna charakteryzacja i organizacja informacji mogą pochodzić z analizy częstości i rozkładu słów w dużych kolekcjach dokumentów tekstowych" [6].

Założenia dla text miningu zostały rozwinięte w 1960 roku, kiedy to zostały zbudowane pierwsze systemy komputerowe przetwarzające nieustrukturyzowany tekst. Do połowy lat 80-tych doświadczenia użytkownika nie zostały dużo ulepszone, gdyż systemy nadal skupiały się na paradygmacie wyszukiwania słów kluczowych.

Don R. Swanson wyartykułował pomysł, że literatura naukowa powinna zostać dostarczona jako naturalne zjawisko wynikające z „badań, korelacji i syntezy". Zestawił poglądy naukowców odnośnie użytkowania informacji z poglądami osób zajmujących się analizą dostarczonych danych – informacji [7]. Dla pracującego naukowca lub inżyniera, czas spędzony na zgromadzenie informacji lub napisanie sprawozdania, często jest traktowany jako czas zmarnowany, który mógłby być zużyty na wytworzenie wniosków, uważanych za coś nowego i bardziej pożytecznego. Osoba zajmująca się analizą dostarczonych informacji, myśli przeciwnie, dla niej bardzo ważna jest dostępna baza rejestrowanej informacji. Nowa wiedza lub skończona notatka pozwala zobaczyć jak pojawienie się ogromnych ilości pojedynczo nieistotnych danych tworzy fragmenty, które niekoniecznie zostały rozpoznane jako spokrewnione z sobą, zaś w całości dają konkretną informację i pozwalają wysunąć daleko idące wnioski. Użycie terminu wyszukiwanie informacji jest nieodpowiednią i bardzo mylącą metaforą. Analityk nieustannie współdziała z częściami zgromadzonych danych jak gdyby były one częściami wybranymi z tysiąca elementów układanki. Poszukiwane są istotne wzorce w na pozór nie związanych ze sobą dokumentach. Swanson zalecał, aby naukowcy wzięli na poważnie pomysł, że nowa informacja, a co za tym idzie wiedza, może zostać uzyskana z kolekcji dokumentów, jak również

rozszerzyć poglądy odnośnie niedostrzegalnej informacji uzyskanej z badań. Prace Swansona, ukierunkowane na odkrycie znaczącej nowej wiedzy w literaturze biomedycznej, przyniosły mu duży sukces. Został ona uznany za wczesnego pioniera text miningu, a w ślad za nim poszli Marti Hearst i Ronald Kostoff [2, 8]. Można posunąć się dalej i zaproponować, że z powodu pomysłów jakie wygłaszał w nieznanej jeszcze dziedzinie eksploracyjnej analizy tekstów, można uznać Swansona ojcem współczesnego text miningu.

Dalszy rozwój narzędzi eksploracyjnej analizy tekstu nastąpił w latach 90-tych, kiedy pojawiło się przetwarzanie języka naturalnego (NLP) oraz sztuczna inteligencja (AI), na których opiera się text mining. Metody rozwinięte w tym okresie są nadal używane w dostępnych dzisiaj narzędziach text miningowych. Dalsze badania naukowe prowadzone w zakresie lingwistyki obliczeniowej (*computational linguistics*) okazały się na tyle owocne, że zaczęto wytwarzać oprogramowanie do text miningu. Badania nad text miningowymi metodami eksploracji danych wydają się być bardzo obiecujące, gdyż pozwalają na zaoszczędzenie czasu i pieniędzy, które musiałyby zostać przeznaczone na przeczytanie i ewentualne eksplorowanie przez człowieka ogromnego repozytorium dokumentów tekstowych.

1.3. Problemy badawcze eksploracyjnej analizy tekstów

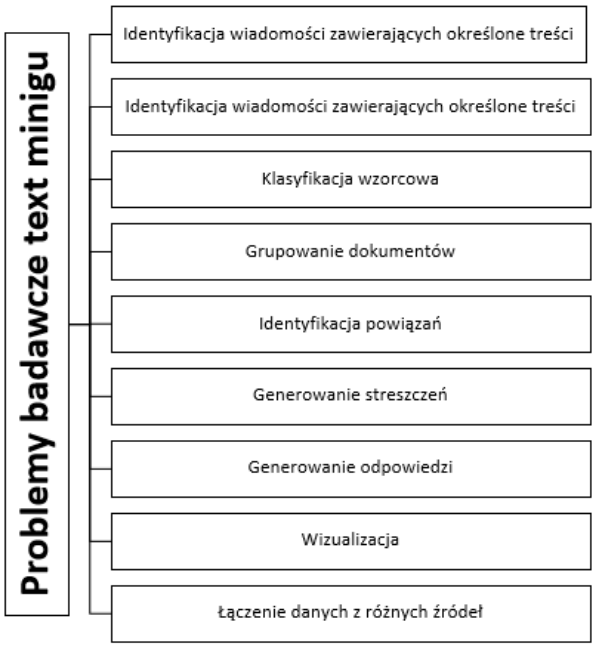
Oczywistym problemem wprowadzenia text miningu jest fakt, że język naturalny został rozwinięty „dla ludzi", by ułatwić komunikację między nimi i rejestrować informację, zaś komputery muszą przejść długą drogę, aby zrozumieć język naturalny. Ludzie posiadają zdolności, by odróżniać i stosować lingwistyczne wzorce tekstowe i mogą łatwo pokonywać przeszkody, z którymi komputery sobie nie radzą, takie jak regionalizm, slang, formy gramatyczne czy też rozumienie kontekstowe. Jednakże, chociaż nasze zdolności językowe pozwalają nam zrozumieć nieustrukturyzowane dane (tekst), brakuje zdolności komputera, by je przetworzyć. Kluczem technologii text miningowej jest łączenie lingwistycznych zdolności człowieka z szybkością i dokładnością komputera.

Text mining ma za zadanie [9]:

- dostarczać teksty związane z danym obszarem zainteresowania (tradycyjne IR),
- opisywać zawartość tekstu w sposób przydatny do dalszego przetwarzania (przetwarzanie języka naturalnego, modelowanie statystyczne itp.),
- interpretować i analizować informację wynikową (znajdując powiązania, interesujące tematy).

W myśli przedstawionych zadań stawianych przed text miningiem można wyodrębnić następujące główne problemy badawcze eksploracyjnej analizy tekstów zobrazowane na Rys. 1.:

- Pozyskiwanie informacji z dokumentów – mechanizmy pozyskiwania bazują przede wszystkim na próbie dopasowania do poszczególnych fragmentów tekstu wzorców określających rodzaj poszukiwanych treści. Podstawowym celem prac w tym zakresie jest stworzenie systemu wspomagającego proces pozyskiwania informacji z literatury danego przedmiotu, nadając im zdefiniowaną strukturę i umieszczając w bazie danych [4,10].
- Identyfikacja wiadomości zawierających określone treści – celem analizy jest stworzenie systemu monitorującego dużą liczbę dokumentów w celu identyfikacji tych, które mogą być istotne z punktu postawionego kryterium. Automatyzacji podlega jedynie wskazanie potencjalnych dokumentów, bez ich analizowania [18].



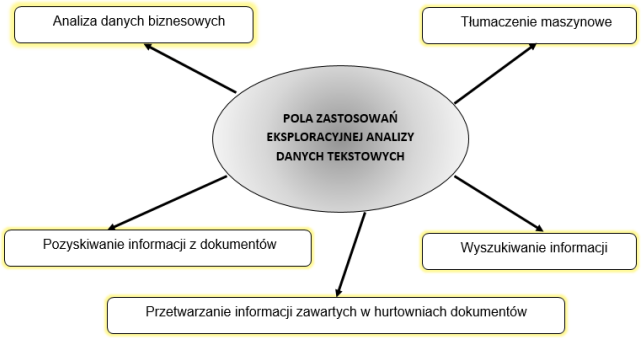
Rys. 1. *Główne problemy badawcze text miningu*

- Klasyfikacja wzorcowa – analiza zbioru dokumentów i przypisanie każdego z nich, po uwzględnieniu informacji w nich zawartych, do jednej z wcześniej wyróżnionych klas. Przyjmuje się, iż jeden dokument może należeć do więcej niż jednej klasy. Badania w tym wymiarze polegają na zdefiniowaniu wzorców poszczególnych klas i określeniu sposobu podobieństwa dokumentu do danego wzorca [10]. Klasyfikacja wzorcowa jest bardzo pomocna podczas wyszukiwania oraz filtrowania informacji.
- Grupowanie dokumentów – polega na próbie połączenia w grupy (klastry) dokumentów na podstawie ich podobieństwa, tak aby dokumenty dotyczące jednego tematu trafiły do tej samej grupy. Innymi słowy problem polega na takim pogrupowaniu dokumentów, by dokumenty należące do jednej klasy były do siebie możliwie podobne i jednocześnie różniły się znacząco od dokumentów należących do innych klas. Najważniejszym zadaniem jest określenie liczbowej miary podobieństwa między dokumentami. Szczegółowe omówienie tematu grupowania danych tekstowych zaprezentowane zostało w [11, 25].
- Identyfikacja powiązań – rozumiana jako wykrycie związków istniejących pomiędzy informacjami pozyskanymi z dokumentów tekstowych lub też jako identyfikacja dokumentów, które są powiązane ze sobą ze względu na zawarte w nich treści [18].
- Wizualizacja – zazwyczaj jest powiązana z próbą rozwiązania innego typu zadania. Bardzo często interpretacji graficznej poddawane są związki zachodzące pomiędzy wyodrębnionymi faktami lub zależności występujące w strukturze analizowanego zbioru dokumentów.
- Sumaryzacja polegająca na generowaniu streszczeń – podstawowym założeniem jest automatyczne przygotowanie konspektu z dowolnego dokumentu podsumowującego jego zawartość [4]. Rozwiązanie tego problemu jest bardzo przydatne do analizy dużych zbiorów tekstowych.
- Generowanie odpowiedzi na pytania - głównym polem badań jest możliwość zrozumienia przez komputer pytania sformułowanego przez człowieka w języku naturalnym. Automatyczna interpretacja pytań i udzielanie odpowiedzi doprowadziła do powstania systemów Question&Answering (Q&A).

– Przeprowadzanie rozszerzonego modelowania predykcyjnego polegającego na łączeniu danych pochodzących z różnych źródeł [12].

2. ANALIZA TEXT MININGOWA – ZASTOSOWANIA

Zakres zastosowań eksploracyjnej analizy tekstów został zobrazowany na Rys. 2.



Rys. 1. *Pola zastosowań text miningu*

2.1. Pozyskiwanie informacji z dokumentów

We współczesnym świecie istnieje wiele różnych branż biznesowych, których pracownicy mają problemy z ogarnięciem olbrzymiej ilości ciągle przybywającej informacji, właściwie jej nie używają, ponieważ nie mają żadnej rozsądnej metody analizowania tych danych, informacji. Narzędzia text miningu mogą pomóc tym branżom analizować ich konkurencję, bazę klienta i strategie marketingu, tym samym przyczyniając się do podwyższenia zysków finansowych firm płynących z zakupu oprogramowania text miningowego. Aby zarządzać z powodzeniem wykorzystując nowy projekt text miningu, przedsiębiorstwa muszą być pewne, co do:

– Wyraźnie zdefiniowanego celu i oczekiwań odnośnie projektu text miningowego. Cel projektu powinien być zgodny ze strategicznym celem i wizją przedsiębiorstwa.
– Przeprowadzenia analizy zwrotu z inwestycji za pomocą wskaźnika ROI (ang. return on investment), aby uzasadnić zarówno namacalne jak i nieuchwytne korzyści dla przedsiębiorstwa. Jasne uzasadnienie kosztu może być konieczne, aby otrzymać konieczne poparcie zarządu dla projektu text mining.
– Przeprowadzenia rozmów z innymi sprzedawcami i ich klientami o ich doświadczeniu rozmieszczenia i wsparcia produktu.
– Zintegrowania projektu text miningowego z istniejącą infrastrukturą informatyczną przedsiębiorstwa. Na przykład niektóre przedsiębiorstwa mogą być w stanie zintegrować oprogramowanie text miningu z ich istniejącym oprogramowaniem magazynującym dane, by dostarczyć bardziej skuteczne wsparcie interesu firmy.
– Wynajęcia profesjonalistów, którzy przeprowadzą szkolenie pracowników firmy z zakresu text miningu. Nowe techniki text miningu są ciągle rozwijane, a co się z tym wiąże na rynku pojawiają się regularnie nowe produkty z zakresu eksploracyjnej analizy danych tekstowych. Przedsiębiorstwa muszą być pewne, że to ich personel jest kształcony i posiada istotną wiedzę, aby w pełni wykorzystać narzędzia text minigu.

Obszarem zastosowań metod pozyskiwania informacji z dokumentów tekstowych są systemy wspierające działalność biznesową. Można wskazać możliwości ich zastosowań w systemach CRM, w których mogą służyć jako mocne narzędzie analizy danych tekstowych dotyczących klientów firmy (poznanie profilu klienta, prognozowanie przejścia klienta do konkurencji, identyfikacja przyczyn odejścia klienta) [13]. Tego typu zastosowania mogą być również pomocne w bankowości, gdzie mogą służyć do analizy korespondencji z klientami banku w celu określenia prawdopodobieństwa niespłacenia zaciągniętego kredytu [14].

2.2. Analiza danych biznesowych

Text mining jest także szeroko wykorzystywany do analizy danych biznesowych. Szybsza analiza informacji prowadzi do lepszej identyfikacji potrzeb klientów, ich przyzwyczajeń i oczekiwań, co z kolei przekłada się na wyniki finansowe. Możliwa staje się automatyczna analiza wypełnionych ankiet o danym produkcie firmy, czy też automatyczna analiza rodzaju opinii (pozytywne, negatywne, neutralne) oraz ich zliczanie [2].

Dzięki zastosowaniu narzędzi eksploracyjnej analizy danych satysfakcja klienta może być mierzona i analizowana już choćby na bazie wypełnionych online kwestionariuszy [15]. Text mining wykorzystywany jest także do automatycznej analizy opisów z kart gwarancyjnych wykrywając typowe i powtarzające się problemy z danymi produktami. W dziale call center firmy możliwa jest identyfikacja języka nadchodzących wiadomości i zgłoszeń.

W aplikacjach BI metody text miningu umożliwiają przeszukiwanie sieci w celu znalezienia ważnych i istotnych informacji biznesowych niezbędnych do podejmowania właściwych decyzji odnośnie współpracy z innymi firmami. Zalety wynikające z zastosowania text miningu to między innymi: dostarczenie bardziej precyzyjnych wyników wyszukiwania, zamiana nieustrukturyzowanych dokumentów na ustrukturyzowane, gotowe do wykorzystania, identyfikacja wiadomości zawierających ściśle określone treści, czy też identyfikacja słów kluczowych. Zgłębianie danych tekstowych pozwala na pełne wykorzystanie posiadanych danych o klientach, czy też transakcjach, umożliwia odkrycie krytycznych informacji, które można przekształcić w przewagę konkurencyjną. Metody text miningu zapewniają możliwość prowadzenia wielowymiarowych analiz danych historycznych, jak i prognozowanie wybranych wskaźników ekonomicznych, przy zaistnieniu określonych warunków w przyszłości [14]. Korzystanie z nowych informacji np. z wniosków pozyskanych z analizy dokumentów w procesie text mining, którymi konkurencja może nie dysponować, służy wypracowywaniu pewnej przewagi. Dzięki trafnym decyzjom podjętym na podstawie wyników analiz danych wzrasta konkurencyjność przedsiębiorstwa, a co za tym idzie, także jego zyski.

Sektor ubezpieczeniowy to branża, w której w sposób naturalny od dawna mocno zakorzeniona jest kultura eksploracyjnej analizy danych. Dane te są wykorzystywane m.in. do szacowania wysokości składki, poziomu rezerw szkodowych itp. Każda firma ubezpieczeniowa zatrudnia w tym celu certyfikowanych aktuariuszy oraz wykorzystuje oprogramowanie statystyczne. Ponadto dane są analizowane na potrzeby wykrywania wyłudzeń odszkodowań (z wykorzystaniem analiz text mining) oraz na potrzeby zarządzania relacjami z klientem (analityczny CRM).

Wolumeny danych zbieranych i przetwarzanych w branży finansowej są zdecydowanie mniejsze niż w branży telekomunikacyjnej, nie zmienia to jednak faktu, że banki także nie od dziś analizują dane o swoich klientach. Początki analizy danych w bankowości dotyczą obszaru oceny ryzyka związanego z klientem (scoring kredytowy) [17]. Początkowo analizowane były jedynie dane z aplikacji klientów (scoring aplikacyjny). Obecnie wykorzystywane są także dane o zachowaniach klientów: klienci regularnie są oceniani pod kątem skłonności do zaprzestania spłaty zobowiązania, jakie stanowią zaciągnięte kredyty (scoring behawioralny) [18]. W bankach modele statystyczne są budowane także na potrzeby nadawania ocen ratingowych przedsiębiorcom. Do pozostałych obszarów wykorzystania zaawansowanej analizy danych w sektorze bankowym, jednak mniej rozpowszechnionych niż scoring kredytowy, należą: analityczny

CRM, wykrywanie nadużyć – przy składaniu wniosków, na transakcjach kartowych, zapobieganie praniu brudnych pieniędzy – AML, oraz scoring windykacyjny.

2.3. Przetwarzanie informacji zawartych w hurtowniach dokumentów

W pracy [15] omówione zostały możliwości zastosowań systemów identyfikacji wiadomości wg określonego kryterium jako podstawowego narzędzia stosowanego przy przetwarzaniu informacji przechowywanych w tekstowych hurtowniach danych. Jednym z problemów związanych z zarządzaniem organizacją jest pozyskiwanie i przechowywanie danych będących źródłem wiedzy. Szacuje się że ponad 80% danych, istotnych dla efektywnego zarządzania organizacją, jest przechowywanych w formie dokumentów tekstowych. Konieczne jest zatem tworzenie centralnych repozytoriów danych nieustrukturyzowanych bądź słaboustrukturyzowanych – hurtowni dokumentów. Hurtownię dokumentów charakteryzują następujące cechy, których spełnienie możliwe jest dzięki narzędziom text miningu [19, 20]:

- Implementacja języka zapytań umożliwiającego wyszukiwanie dokumentów na podstawie ich atrybutów i słów kluczowych;
- Gromadzenie i udostępnianie metadanych opisujących poszczególne dokumenty;
- Pobieranie i przechowywanie istotnych cech każdego dokumentu (słowa kluczowe, streszczenia, indeksy) niezależnie od samego dokumentu;
- Automatyczna klasyfikacja dokumentów na podstawie kryteriów definiowanych przez użytkownika;
- Możliwość automatycznego grupowania dokumentów;
- Przechowywanie informacji o semantycznych powiązaniach pomiędzy dokumentami.

Dzięki zgłębieniu danych zawartych w bazach przedsiębiorstwa można m.in. określić:

- jacy klienci opuszczają przedsiębiorstwo,
- jakie jest ryzyko odejścia poszczególnych klientów,
- cechy klientów gotowych opuścić przedsiębiorstwo,
- charakterystykę grup klientów,
- jakie sprzężone produkty są nabywane przez poszczególne grupy klientów (analiza koszykowa).

Mając taki zestaw informacji, wzbogacony o analizę satysfakcji i preferencji, przedsiębiorstwo może konstruować skuteczne programy lojalnościowe.

2.4. Gospodarka elektroniczna

Konsekwencją ciągłego rozwoju i upowszechniania się technologii informacyjnej, w tym także dynamicznego rozwoju sieci globalnych, jest powstanie gospodarki elektronicznej. Jest to obecnie zjawisko o coraz większym znaczeniu gospodarczym, zmieniające istotę prowadzenia działalności gospodarczej. Gospodarka elektroniczna jest także określana jako gospodarka cyfrowa, e-gospodarka lub nowa gospodarka, stanowi nowy paradygmat biznesu. Można ją określić jako wirtualną arenę, na której jest podejmowana działalność, przeprowadzane są transakcje, dochodzi do tworzenia i wymiany wartości oraz powstają i zacieśniają się bezpośrednie kontakty między jego uczestnikami [21 s. 69]. Istotnym paradygmatem e-przedsiębiorstwa jest wykorzystywanie dokumentów elektronicznych, czyli zastąpienie dokumentów papierowych dokumentami elektronicznymi i wykorzystanie poczty elektronicznej do obrotu nimi. Elektroniczne dokumenty łatwo jest rozpowszechniać, gdyż znika bariera odległości, skraca się czas ich przesyłania, a koszty są minimalne. Coraz częściej najważniejszym warunkiem osiągnięcia wysokiej jakości informacji jest stosowanie zasady selekcji. W gospodarce elektronicznej menedżer nie odczuwa braku informacji, a wręcz przeciwnie, cierpi na jej nadmiar. Konieczne jest zatem selekcjonowanie informacji, do którego z powodzeniem można zastosować metody analizy text miningowej.

Do sprawnego funkcjonowania e-przedsiębiorstwa szczególnie potrzebna jest kompleksowa obsługa informacyjna procesów decyzyjnych. W dzisiejszych czasach dobre zarządzanie biznesem to zarządzanie jego przyszłością, a więc zarządzanie informacją [22, s. 38].

W wielu e-przedsiębiorstwach informacja związana z klientem jest rozproszona w kilku miejscach i ma różny format, co utrudnia jej integrację i przetwarzanie. Przy zwiększającej się ilości informacji niezbędne jest współdzielenie informacji o kliencie i uzupełnianie w czasie rzeczywistym danych marketingowych. W działaniach tych często wykorzystuje się technologię informatyczną wykorzystującą eksploracyjną analizę danych tekstowych, w tym tak zwane systemy zarządzania relacjami z klientami (*Customer Relationship Management*). Głównym ich celem jest bezpieczeństwo, komfort i szybkość przesyłania informacji między e-przedsiębiorstwem a jego klientem, co gwarantuje, że zostanie on obsłużony w profesjonalny sposób. Wdrożenie systemu CRM umożliwia redukcję kosztów funkcjonowania działów obsługi klienta i jednocześnie zwiększa ich skuteczność.

2.5. Wyszukiwanie informacji

Rozwój społeczeństwa informacyjnego oraz technologii informatycznych pociągnął za sobą w sposób naturalny powstanie zautomatyzowanych systemów wspomagających wyszukiwanie i porządkowanie informacji. W obecnych czasach, kiedy dostęp do Internetu jest praktycznie nieograniczony a ilość zgromadzonych tam informacji ogromna, bezdyskusyjnym staje się potrzeba szybkiego i efektywnego wyszukiwania w zasobach sieci potrzebnych informacji. Różnorodne narzędzia IR realizujące to zadanie powstawały niemal równolegle z upowszechnianiem się Internetu. Klasyczne struktury danych występujące w analizie text miningowej (np. TF-IDF, SVD) zostały szeroko wykorzystane w zadaniach IR [23], stanowiąc niejako uzupełnienie procesu wyszukiwania informacji.

2.6. Tłumaczenie maszynowe

Tłumaczenia maszynowe są nieodzownym elementem w dziedzinie tłumaczeń, szczególnie w czasach szybko rozwijających się technologii komputerowych. Właściwie zastosowane, mogą stanowić fantastyczne narzędzie pomagające w pracy nad przekładem, przede wszystkim przyspieszając i usprawniając cały proces tłumaczenia. Warto jednak pamiętać, że rzadko kiedy tekst przetłumaczony maszynowo będzie gotowym produktem, który można bez korekty przekazać odbiorcy. Większość tłumaczeń wygenerowanych przez rozmaite programy jest zaledwie tłumaczeniem wstępnym, wymagającym obróbki, m.in. w zakresie stylistyki, szyku zdań, czy gramatyki. Mimo, że tłumacz musi wykonywać jeszcze wiele pracy nad maszynowo wygenerowanym tekstem, tłumaczenia tego typu znajdują rozmaite zastosowanie w dziedzinie przekładów.

Dla wielu firm nadrzędnym celem strategii marketingowej jest zdobycie nowych rynków wraz z ich potencjałem konsumenckim. O sukcesie bądź fiasku decydują wówczas w dużym stopniu czynniki językowe i kulturowe, jednak zagadnienia te bywają często zaniedbywane i traktowane bardzo powierzchownie. Przedsiębiorcy chcą działać szybko i tanio, produkt musi znaleźć się na nowym rynku „na wczoraj". Jedynym wyzwaniem staje się tłumaczenie — im szybsze i tańsze, tym lepiej. Mało które firmy dostrzegają korelację między właściwym przygotowaniem innej wersji językowej produktu a jego późniejszym sukcesem lub fiaskiem na nowym rynku. Tłumaczenia są często traktowane jako koszt, tym bardziej że ogromna część tekstów jest do siebie podobna, a czasem wręcz tłumaczy się wielokrotnie

identyczne fragmenty. Tę prawidłowość wykorzystuje rozwijana technologia wspomagania tłumaczeń (*CAT — Computer-Aided Translation*) [24]. Stosowanie narzędzi CAT umożliwia odejście od archaicznych metod wyceny (znaki, strony) na rzecz realistycznej analizy dokumentów (wykorzystując eksploracyjną analizę) jeszcze przed wykonaniem tłumaczenia.

PODSUMOWANIE

Nadmiar dostępnych danych okazuje się trudny do ogarnięcia i analizy, co stanowi poważny o ile nie najpoważniejszy problem pojawiający się na styku człowiek — komputer. Człowiek dysponuje umysłem daleko doskonalszym od najlepszej maszyny, jednak jego możliwości percepcji są mocno ograniczone. Komputer, pomimo dużej objętości pamięci i szybkości działania potrafi w znikomym stopniu zrozumieć człowieka i świat (odgadnąć jego zamiary, intencje, czy potrzeby). Nie dziwi zatem, że to właśnie rozumienie języka naturalnego przez maszyny wydaje się być najbardziej ambitnym, a zarazem najodleglejszym celem jaki może osiągnąć informatyka.

Z uwagi na to, że ostatecznym odbiorcą danych i informacji jest człowiek, dla niego dokonuje się wszelkiego rodzaju przetwarzania danych. Problemem, bowiem nie jest efektywne gromadzenie i przetwarzanie danych, lecz zdolność interpretacji i wyciągania użytecznych wniosków [26].

Rozwój społeczeństwa informacyjnego oraz technologii informatycznych pociągnął za sobą w sposób naturalny powstanie zautomatyzowanych systemów wspomagających wyszukiwanie i porządkowanie informacji. Techniki text miningu znajdują coraz większe zastosowanie w ekonomii, medycynie, lingwistyce, biznesie oraz innych aspektach zarządzania, w których występuje problem nadmiaru informacji zapisanej w języku naturalnym.

BIBLIOGRAFIA

1. Kao A., Poteet S., Natural Language Processing and Text Mining, Springer, 2007.
2. Hearst M., *Untangling Text Data Mining*, Proceedings of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics, University of Maryland, June 20-26 1999.
3. Fan W., Wallace L., Rich S., *Tapping into the Power of Text Mining*, Communications of ACM, 2005.
4. Liddy E. D., *Text mining*, Bulletin of the American Society for Information Science 27, 2000.
5. Luhn H. P., *The automatic creation of literature abstracts*, IBM Journal of Reasearch and Development, 1958, s. 159–165.
6. Doyle L, B., *Semantic Road Maps for Literature Searchers*, The Journal of the Association of Computing Machinery(ACM), Vol, 8(4), 1961, s. 553-578.
7. Swanson D, R., *Complementary structures in disjoint scientific literatures*, Proceedings of the Fourteenth Annual International ACM SIGIR Conference on Research and Development, ACM Press, New York 1991, s. 280-289.
8. Kostoff R. N., *Science and technology innovation*, Technovation 19(10), 1999, s. 593-604.
9. Perrin, P., Personal communication, Molecular Systems research group, Merck & Co., Inc., Rahway, New York, 2001
10. Hand D., Mannila H., Smyth P., *Eksploracja danych*, Wydawnictwo Naukowo-Techniczne, Warszawa, 2005.
11. Larose D. T., *Odkrywanie wiedzy z danych*, Wydawnictwo Naukowe PWN, Warszawa, 2006.
12. Berry M. W., Kogan J., *Text Mining Applications and Theory*, John Wiley & Sons Ltd, United Kingdom, 2010.
13. Shull S., *Do You Know What Your Customers are Telling You?*, DM Direct Special Report, May 10, 2005.
14. Brown P. F., Cocke J., Della Pietra S. A., Della Pietra V. J., Jelinek F., Lafferty J. D., Mercer R. L., Roossin P. S., *A statistical approach to machine translation*, Computational Linguistics Vol, 16, no 2, 1990.
15. Night K., Mining Online Text, Communications of the ACM 42(11), ACM Press, New York, 1999, s. 58–61.
16. Billewicz A., *Budowa procesów ekstrakcji, transformacji i ładowania danych w systemach Business Intelligence*, (red,) Sroka H., Porębska T., Systemy wspomagania organizacji SWO2004, Katowice, 2004.
17. Janc A., Kraska M., *Credit-scoring: nowoczesna metoda oceny zdolności kredytowej*, Biblioteka Menedżera i Bankowca, Warszawa, 2001, s. 14-15.
18. Wójciak M., *Metody oceny ryzyka kredytowego*, PWE, Warszawa 2007.
19. Sullivan D., Document Warehousing and Text Mining: Techniques for Improving Business Operations, Marketing and Sales, Wiley Computer Publishing, 2001.
20. Ishikava H., Ohta M., Kato K., *Document Warehousing: A Document-Intensive Application of Multimedia Database*, 11th International Workshop on research Issues in Data engineering, Heidelberg, 2001.
21. Hartman A., Sifonis J., Kador J., E-biznes, Strategie sukcesu w gospodarce internetowej, Sprawdzone metody organizacji przedsięwzięć e-biznesowych, Wydawnictwo K,E, Liber s,c., Warszawa, 2001.
22. Kotler P., *Marketing, Analiza, planowanie, wdrażanie i kontrola*, wyd, VI, Gebethner i Ska, Warszawa, 1994.
23. Eden L., Matrix Methods In Data Mining and Pattern Recognition, SIAM 2007.
24. Bowker L., Computer-Aided Translation Technology, A Practical Introduction, University of Ottawa Press, Ottawa, 2002.
25. Manning C. D., Raghavan P., Schütze H., *Introduction to Information Retrieval*, Cambridge University Press, Cambridge, England, 2008.
26. Fayyad U. M., Piatetsky-Shapiro G., Smyth P., *From data mining to knowledge discovery in databases*, AI Magazine, Vol 17, 1996, s. 37-54.

Overview of uses text mining analysis

The article discussed the text mining with particular emphasis on the use of text mining analysis. In the modern world there are many different business industries where workers are in contact with an excess of incoming information. The development of the information society and information technology entailed a natural rise of automated systems to support search and organize information. Text mining techniques are increasingly applied, and a broad overview of applications, together with an indication of the practical possibilities of the use of text mining analysis has been thoroughly discussed in the article.

Autorzy:
dr inż. **Anna Gładysz** – Politechnika Rzeszowska