ZESZYTY NAUKOWE POLITECHNIKI RZESZOWSKIEJ Nr 280
Zarządzanie i Marketing z. 18 (3/2011) 2011

Anna GŁADYSZ[1]

PRZETWARZANIE INFORMACJI – WYBRANE ALGORYTMY GRUPOWANIA DANYCH W KOLEKCJI DOKUMENTÓW TEKSTOWYCH

Zasadniczym celem referatu jest zaprezentowanie problemów związanych z grupowaniem danych w kolekcji dokumentów tekstowych. Przedstawiono przegląd i klasyfikację algorytmów grupowania dokumentów tekstowych. Metody grupowania zebrane zostały w kilka kategorii uzależnionych od ogólnego mechanizmu działania: metody płaskie, hierarchiczne, grafowe i inne. Dla stworzonej hierarchii grup dokumentów tekstowych, z każdej grupy należy wydobyć słowa kluczowe. W tym celu wykorzystane zostały metody stosowane na gruncie eksploracyjnej analizy dokumentów tekstowych (text mining).

1. WPROWADZENIE

Dokumenty tekstowe są jedną z najpopularniejszych i najczęściej spotykanych form zapisu informacji. Często są źródłem ważnej i użytecznej wiedzy. Jednak ten typ danych trudno jest analizować z powodu niskiego stopnia ustrukturyzowania tekstu, wieloznaczności wypowiedzi czy też braku jednoznacznych metod interpretacji tekstów. W odniesieniu do dokumentów testowych możemy zastosować metody eksploracji danych. Z praktycznego punktu widzenia szczególnie przydatne są dwie metody eksploracji tekstu: klasyfikacja i grupowanie. W dalszej części artykułu opisane zostaną problemy i wybrane algorytmy grupowania danych w kolekcji dokumentów tekstowych.

2. GRUPOWANIE DANYCH

Grupowanie danych jest koncepcją eksploracji danych polegającą na przeszukiwaniu i łączeniu (grupowaniu) danych wewnątrz zbioru w klastry często nazywane skupieniami. Wewnątrz klastrów obiekty powinny być do siebie jak najbardziej podobne – „bliskie", natomiast obiekty różnych klastrów najbardziej różne od siebie – „odległe"[Koronacki i Ćwik 2005]. Grupowanie może dotyczyć zarówno obiektów rzeczywistych (np. pacjentów, sekwencji DNA, dokumenty tekstowe), jak również obiektów abstrakcyjnych (sekwencja dostępów do stron WWW, grafy reprezentujące dokumenty XML, itp.).

Zagadnienie grupowania (klasteryzacji) polega na podziale zadanego zbioru dokumentów tekstowych na pewną liczbę grup (klastrów), w ramach, których dokumenty charakteryzują się podobną treścią. Podobieństwo rozumiane jest najczęściej, jako podobieństwo tematyki dokumentów, ale możliwe jest także grupowanie według innych kryteriów, np. według pewnych cech stylu. Pożądane jest, aby dokumenty przydzielone do tej samej grupy były do siebie wzajemnie jak najbardziej podobne, zaś dokumenty przydzielone do różnych grup powinny się między sobą jak najbardziej różnić.

[1] Mgr inż. Anna Gładysz, Zakład Informatyki w Zarządzaniu, Wydział Zarządzania, Politechnika Rzeszowska.

Pojawia się problem jak reprezentować dokument tekstowy, aby można było określić „bliskość" pomiędzy dwoma dokumentami. Rozwiązanie może być tak dobre, jak stworzony model problemu.

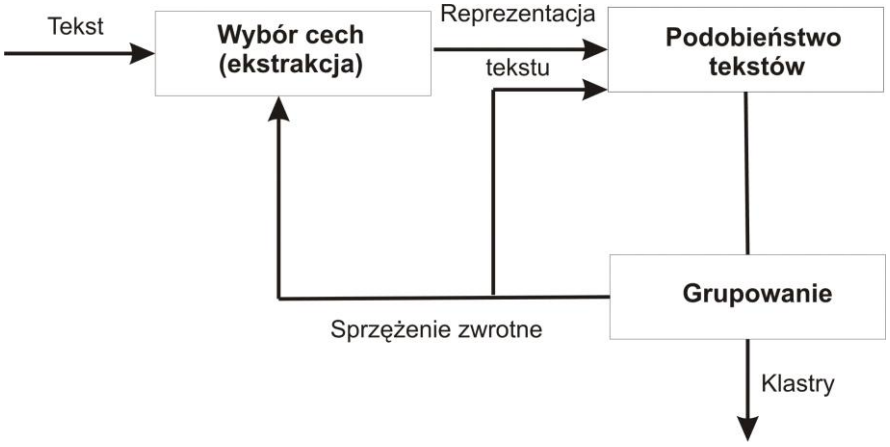
W przypadku, gdy model nie przechowuje informacji np. o nagłówkach sekcji metoda grupowania będzie traktować wszystkie słowa w dokumencie, jako równoważne – co może wpłynąć na jakość grupowania. Jednak z drugiej strony przechowywanie takich informacji może w znacznym stopniu wydłużyć czas działania algorytmu i jego nieużyteczność dla dużych kolekcji dokumentów.

2.1. Składowe procesu grupowania

Proces grupowania jest procesem wieloetapowym i interakcyjnym. Można wyróżnić następujące składowe procesu grupowania (Rys. 1):

I. Wybór cech (atrybutów) najlepiej charakteryzujących dany typ obiektu – w wyniku otrzymujemy pewną abstrakcyjną reprezentację dokumentów.
II. Określenie miary podobieństwa pomiędzy grupowanymi obiektami.
III. Wybór metody grupowania, zależnej od reprezentacji obiektów oraz konkretnego algorytmu grupowania.
IV. Analiza otrzymanych klastrów i próba znalezienia ogólnej charakterystyki klastrów.

Rys. 1. Proces grupowania – składowe.



Źródło: opracowanie własne.

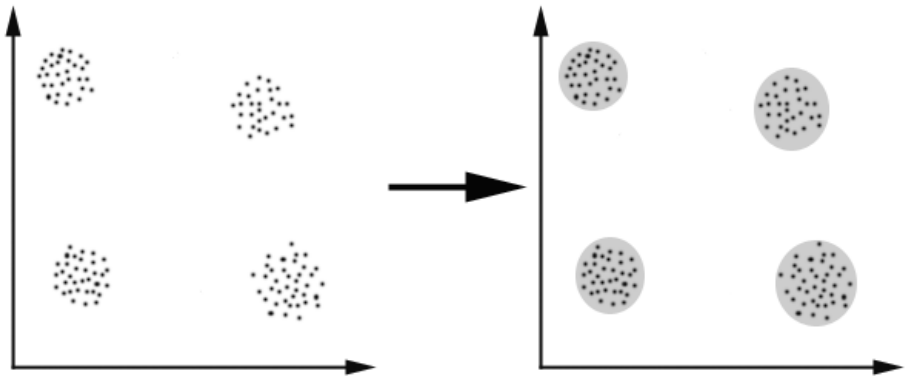
Istnieje szereg wyzwań i problemów związanych z mechanizmami grupowania danych:

- Problem złożoności obliczeniowej podczas grupowania dużej ilości danych, zwłaszcza z przestrzeni wielowymiarowej.
- Efektywność metody zależy od definicji odległości pomiędzy obiektami.
- W przypadku braku standardowej miary mogącej stanowić odległość pomiędzy obiektami, należy zdefiniować własną miarę odległości, co może być trudne i problematyczne, zwłaszcza w wielowymiarowej przestrzeni cech.
- Problem interpretacji wyników działania algorytmu grupowania danych.

Do interpretacji wyników działania algorytmu grupowania danych często wykorzystywane są narzędzia umożliwiające wizualizacje wyników (grup i przynależnych im obiektów). Bardzo prosta wizualizacja i zarazem schemat działania algorytmu grupowania przedstawiony jest

na rysunku (Rys. 2). Jednak wizualizacja graficzna sprawdza się w przestrzeniach co najwyżej trójwymiarowych. W przypadku przestrzeni powyżej trzech wymiarów prosta wizualizacja staje się nieczytelna.

Rys. 2. Ogólny schemat przedstawiający działania algorytmu grupowania danych.



Źródło: http: //home.dei.polimi.it/matteucc/Clustering/tutorial_html/

3. WYBRANE ALGORYTMY GRUPOWANIA DANYCH

Istnieje wiele różnych metod i algorytmów grupowania a także ich klasyfikacji uwzględniających różne aspekty procesu grupowania. Grupowanym typem danych są dane tekstowe. Ze względu na ten typ danych metody grupowania dzielimy na:

- płaskie – dzielą zbiór danych na części bez wskazywania połączeń pomiędzy grupami: metody k-średnich, k-mediana,
- hierarchiczne – tworzą hierarchię grup,
- oparte na gęstości danych – definiują problem grupowania używając pojęcia gęstości
- grafowe – wykorzystują teorię grafów do budowy grup,
- bazujące na naturze – opierają się na mechanizmach które można zaobserwować w naturze,
- inne – niepasujące do żadnej z wyżej wymienionych kategorii.

Wyodrębnia się także podział na metody twarde (ang. crisp, hard), które charakteryzują sie tym, że jeden obiekt może należeć tylko do jednej grupy i metody miękkie (ang. fuzzy, soft), pozwalające na przynależność jednego obiektu do wielu grup..

3.1. Algorytm grupowania k-średnich

Klasycznymi płaskimi metodami grupowania są: metoda k–średnich (ang. *k–means*) i metoda k–mediana (ang. *k–medoid*). Schemat algorytmu k-średnich został przedstawiony na poniższym rysunku (Rys. 3).

Centroid to obiekt będący średnią z elementów należących do grupy — najczęściej jest to obiekt tworzony sztucznie. Standardowo jako funkcji odległości używa sie odległości euklidesowej (tzw. norma L2) – wzór (1) gdzie w przestrzeni k-wymiarowej obliczana jest odległość

między dwoma punktami $x=[x_1,x_2,\ldots x_k]$oraz $y=[y_1,y_2,\ldots y_k]$ lub odległości Manhattan (tzw. Miara L1) – wzór (2).

$$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \tag{1}$$

$$\sum_{i=1}^{k}|x_i - y_i| \tag{2}$$

W modelu wektorowym lepszym rozwiązaniem jest użyć miary odległości bazującej na kosinusie kąta pomiędzy wektorami [Manning i Schütze 2001].

Rys. 3. Algorytm grupowania k-średnich.

1. Dokonaj podziału wszystkich dokumentów na k grup.
2. Oblicz centroid dla każdej grupy.
3. Porównaj wektor każdego dokumentu z centroidem grupy i zanotuj centroid który jest najbardziej podobny.
4. Przesuń wszystkie dokumenty do ich najbardziej podobnych grup.
5. Jeżeli żadne dokumenty nie został przesunięte do nowej grupy, wtedy koniec; w przeciwnym razie wróć do kroku 2.

Źródło: Na podstawie [Weiss, Indurkhya, Zhang i Damerau, 2005 s. 110].

Jako wejście algorytm przyjmuje liczbę dodatnią k, która określa pożądaną liczbę grup. W pierwszym kroku działania należy wylosować k dokumentów, które stanowią początkowe centroidy. Następnie do każdego centroidu zostają przypisane wszystkie najbliższe mu dokumenty, tworząc grupę. Dla każdej grupy wyznaczany jest punkt środkowy – w ten sposób powstaje nowy centroid. Uaktualnianie środków grup i przypisywanie do nich nowych dokumentów jest powtarzane dopóki chociaż jeden centroid został zmieniony.

Metoda k–średnich wprowadza szereg problemów, które należy rozwiązać. W jaki sposób traktować remisy, czyli sytuacje gdy dokumenty są w takiej samej odległości od więcej niż jednego środka grupy. Należy także z góry wiedzieć ile chcemy mieć grup – w dziedzinie dokumentów tekstowych w większości przypadków taka informacja nie jest wiadoma. Losowy wybór punktów początkowych może prowadzić do złych wyników. Często znajdowane są tylko ekstrema lokalne. Ponadto wykrywa ona jedynie grupy o sferycznych kształtach. Nie jest ona także odporna na szum (rozumiany jako losowość dystrybucji słów lub jako cecha, która zwiększa poziom błędu po uwzględnieniu jej w modelu [Manning, Raghavan i Schütze 2007]) i punkty odstające ([Ng i Han, 1994] definiują punkt odstający jako „punkt danych bardzo odstający od pozostałych" ang. *data points that are very far away from the rest of the data points*). Jednak prostota i duża wydajność uczyniła go popularnym mimo licznych wad.

3.2. Grupowanie hierarchiczne

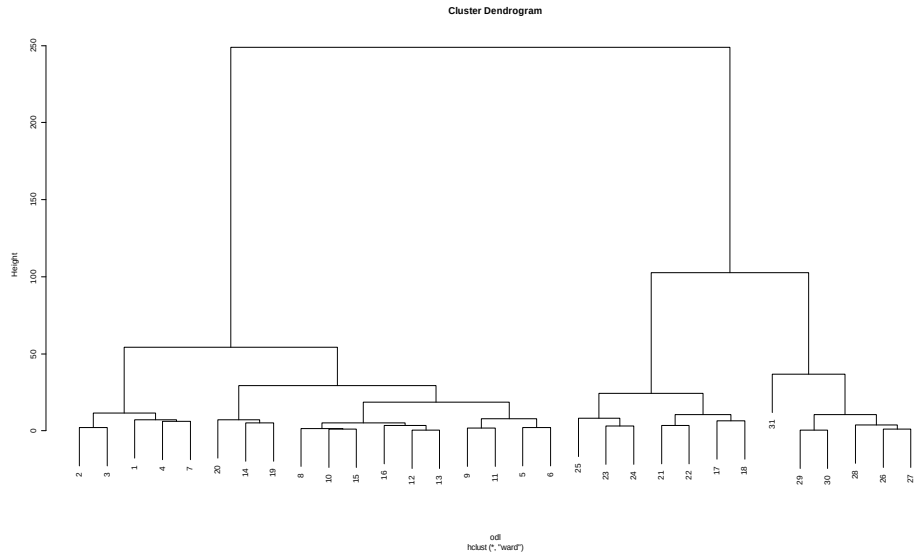
Metody grupowania hierarchicznego polegają na sekwencyjnym grupowaniu obiektów. Sekwencja operacji grupowania tworzy drzewo klastrów, nazywane dendrogramem.

Drzewo grup może być stworzone na dwa sposoby: od dołu(ang. *bottom–up*) poprzez rozpoczęcie od grup zawierających pojedyncze obiekty, w każdym kroku algorytm łączy najbardziej bliskie grupy; od góry (ang. *top–down*) startuje z jedną grupą zawierającą wszystkie obiekty, która jest dzielona w taki sposób, żeby maksymalizować podobieństwo wewnątrz grup. Metoda od dołu nazwana jest aglomeracyjną (HAC, ang. *Hierarchical Agglomerative Clustering*). Drugie podejście nazywane jest deglomeracyjnym lub podziałowym.

Dendrogram przedstawiony na rysunku (Rys. 4) ilustruje działanie hierarchicznego aglomeracyjnego algorytmu grupowania.

Początkowo, wszystkie obiekty 1,2, ... 27 należą do osobnych klastrów. Następnie, w kolejnych krokach, klastry są łączone w większe klastry (łączymy 2 i 3, 4 i 7, 14 i 19, itd. następnie, 1 łączymy z klastrem zawierającym obiekty 4 i 7, zaś 20 łączymy z klastrem zawierającym obiekty 14 i 19 itd.). Proces łączenia klastrów jest kontynuowany tak długo, aż liczba uzyskanych klastrów nie osiągnie zadanej liczby klastrów.

Rys. 4. Przykładowy dendrogram uzyskany na podstawie badań przeprowadzonych w języku R.

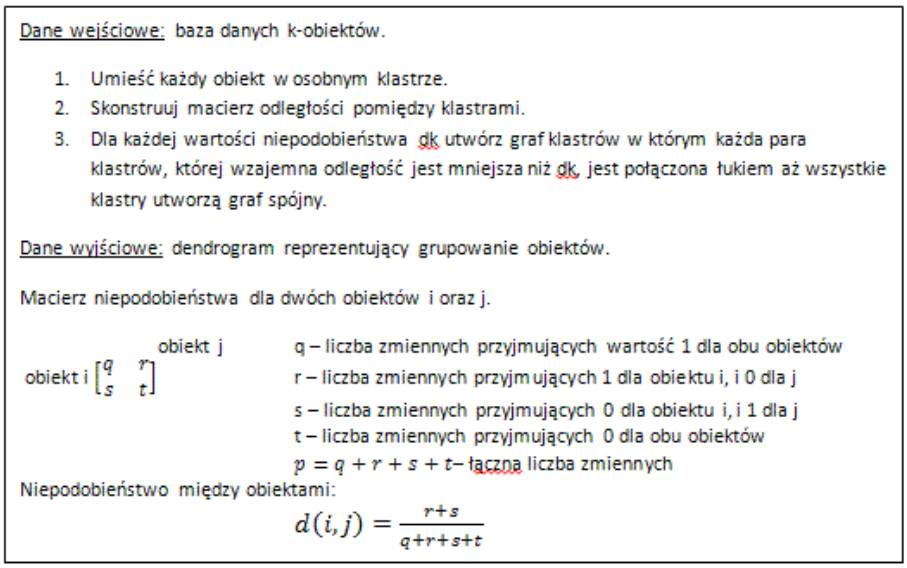


Źródło: opracowanie własne.

Ogólny schemat hierarchicznego aglomeracyjnego algorytmu grupowania można zdefiniować w trzech krokach przedstawionych na rysunku (Rys. 5).

W porównaniu z metodami płaskimi, metody hierarchiczne generują grupy o lepszej jakości, dostarczają większej ilości informacji, jednak są bardziej złożone obliczeniowo.

Rys. 5. Schemat hierarchicznego aglomeracyjnego algorytmu grupowania.



Źródło: opracowanie własne.

3.3. Algorytm DBSCAN

Kolejny algorytm DBSCAN opisany w [Ester, Kriegel, Sander i Xu, 1996] znajduje klastry w oparciu o gęstość danych. Rozpoczyna on działanie od losowo wybranego punktu p. Jeżeli punkt ten jest punktem wewnętrznym, tworzony jest składający się tylko z tego punktu klaster, a następnie dołączane są do niego wszystkie punkty osiągalne z p. Po wyczerpaniu wszystkich takich punktów, wybierany jest losowo pewien nieodwiedzony jeszcze punkt i rozpoczynana jest kolejna iteracja. Jeżeli p jest punktem brzegowym, nie jest on dołączany do żadnego klastra. W odróżnieniu od algorytmów wykorzystujących metody płaskie DBSCAN traktuje niektóre punkty jako szum, który nie powinien zostać przypisany do żadnej ze znajdowanych grup. Algorytm kończy działanie gdy wszystkie punkty zostały już odwiedzone.

Parametrami wpływającymi na sposób działania DBSCAN są: promień sąsiedztwa uwzględnianego przy obliczaniu gęstości danych oraz minimalna gęstość danych potrzebna do uznania punktu za punkt wewnętrzny. Parametry te mogą być ustalone z góry lub mieć przypisane pewne wartości dopasowywane automatycznie do grupowanego zbioru punktów. Mogą one również zostać dobrane na podstawie pewnych cech klastrów znalezionych we wstępnym przebiegu algorytmu DBSCAN, po którym dopiero jest wykonywany przebieg właściwy, korzystający już z dopasowanych parametrów.

Algorytmy oparte na gęstości są rzadko stosowane w dziedzinie dokumentów tekstowych. Wyniki eksperymentów omówione w [Deepak i Roy, 2006] sugerują, że dokumenty naturalnie tworzą grupy wypukłe. Zostały one przeprowadzone jednak tylko na jednym korpusie tekstów dla języka angielskiego.

4. OBRÓBKA DANYCH TEKSTOWYCH

Aby dokonać i przedstawić wyniki działania wybranych procedur grupowania należy dane tekstowe wstępnie przetworzyć. Po wczytaniu analizowanych danych do platformy programistycznej wyposażonej w interpretator języka R następuje odczyt plików źródło-wych i utworzenie kolekcji dokumentów za pomocą funkcji *Corpus()*. Zastosowanie trans-formacji przekształcających każdy dokument w kolekcji na małe litery, zastosowanie stop-listy[2] oraz stemmingu[3] (wykorzystując algorytm Portera). Utworzenie macierzy częstości[4] oraz dalsze jej przetwarzanie.

Przeprowadzona analiza została wykonana dla losowego zbioru dokumentów tekstowych w formacie txt. Zbiór ten zawiera 55 dokumentów tekstowych anglojęzycznych o pięciu kategoriach. Kategoriami są: medycyna (10), elektronika (8), optyka (10), biologia (8), pozostałe (19). Zestaw powstał jako losowo wybrane z sieci Internet tytuły i abstrakty zapisane w formacie plików tekstowych. Wykonane analizy zostały przeprowadzone w języku R z wykorzystaniem pakietów *tm*, *cluster, stats* i *clv*.

Wykorzystując metodę k-średnich nie uzyskano dobrych wyników. W tej metodzie podajemy na ile klastrów chcemy podzielić dany zbiór. Dla tego zbioru danych liczba siedmiu klastrów po jednym użyciu nie jest najlepszym dopasowaniem. Zalecane jest uruchomienie kilkakrotne tego algorytmu oraz wybranie najlepszego podziału na klastry.

Najlepszą metodą dla zadanej próbki danych okazała się metoda *complete*. Inną metodę klastrowania hierarchicznego ukazuje funkcja *hclust()*, której wywołanie wraz z metodą *ward* daje pożądane efekty.

Poza przedstawionymi metodami grupowania w pakiecie R dostępnych jest wiele innych metod do analizy skupień: metoda hierarchicznej analizy skupień przez dzielenie czy metoda klastrowania rozmytego. W zależności od testowanego zbioru danych tekstowych należy dobierać odpowiednie metody. Wyniki w przypadku wybranej próbki danych tek-stowych jednoznacznie wskazują na lepsze wykorzystanie metod hierarchicznych.

5. PODSUMOWANIE

Grupowanie dokumentów tekstowych jest bardzo istotnym zagadnieniem. Przykładem zastosowania może być grupowanie przychodzących wiadomości mailowych. Zbiór wiadomości (dokumentów) można zinterpretować jako zbiór punktów w przestrzeni wielowymiarowej, w której pojedynczy wymiar odpowiada jednemu słowu. Współrzędne dokumentu można zdefiniować względną częstością występowania słów. Klastry dokumentów odpowiadają zaś, grupom dokumentów dotyczących podobnej tematyki. Dobór odpowiednich metod grupowania dokumentów tekstowych ma bardzo duży wpływ na dalszą analizę tekstu.

[2] Stop-lista to lista odrzucanych słów. Eliminacja słów o małym znaczeniu, oraz słów popularnym w danym języku, które nie wpływają na identyfikację dokumentu.

[3] Stemming – redukcja do rdzenia. Polega na sprowadzeniu słów do ich form podstawowych, czyli zastąpienia wyrazu ich rdzeniem słowotwórczym.

[4] W reprezentacji wektorowej dowolny dokument jest reprezentowany w postaci wektora częstości występowania słów kluczowych. Stąd zbiór N przechowywanych dokumentów tekstowych można przedstawić w postaci macierzy częstości, gdzie wiersze to wyrazy zaś kolumny reprezentują dokumenty.

LITERATURA

[1] Deepak, P., Roy, S., *Optics on text data: Experiments and test results*, Raport tech., IBM India Research Lab, 2006
[2] Ester M., Kriegel H., Sander J., Xu X., A *Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*, Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96), 1996
[3] Kang N., Domeniconi C., Barbara D., *Categorization and Keyword Identification of Unlabeled Documents*, Fifth IEEE International Conference on Data Mining, Houston, Texas, November 27-30, 2005
[4] Kłopotek M. A., *Inteligentne wyszukiwarki internetowe*, Akademicka Oficyna Wydawnicza EXIT, Warszawa 2001
[5] Koronacki, J., Ćwik, J., *Statystyczne systemy uczące się*, Wydawnictwo Naukowo-Techniczne, Warszawa, 2005
[6] Manning, C. D., Raghavan, P., Schütze, H., *Introduction to Information Retrieval*, Cambridge University Press, 2007
[7] Manning, C. D., Schütze, H., *Foundations of Statistical Natural Language Processing*, The MIT Press, 2001
[8] Ng, R. T., Han, J., *Efficient and effective clustering methods for spatial data Mining*, W VLDB'94, Proceedings of 20th International Conference on Very Large Data Bases, September 12-15, Santiago de Chile 1994
[9] Weiss S.M., Indurkhya N., Zhang T., Damerau F. J., *Text Mining Predictive Methods for Analyzing Unstructured Information*, Springer Science+Business Media Inc., New York USA 2005

PROCESSING INFORMATION – SELECTED ALGORITHMS OF DATA GROUPING IN THE COLLECTION OF TEXT DOCUMENTS

The main aim of the paper is to present the problems associated with grouping of data in the collection of text documents. It was presented an overview and classification of algorithms in grouping for text documents. Grouping methods have been collected in several categories dependent on the overall mechanism of action: methods of flat, hierarchical, graph, and others. For the created hierarchy of text documents, from each group the keywords should be selected. For this purpose, there have been used the methods applied on the basis of exploratory analysis of text documents (text mining).