FINDING SIMILAR DOCUMENTS IN WEB SEARCH RESULTS

Urszula Kużelewska

Faculty of Computer Science, Bialystok University of Technology, Białystok, Poland

Abstract: Searching the Web is a challenging task. According to the Zamir and Etzioni's definition, Internet is "unorganized, unstructured and decentralized place". Although there are powerful search engines available, the number of indexed web pages exceeds 1 trillion [**?**] and still grows. Most of the search engines return list of documents from their bases sorted according to their relevance to a search query. Such approach is not the best, because the returned list is very long and may contain documents not related to the query. To increase efficiency of a searching process one may identify groups of similar documents from result list. One of the tools to do it are traditional clustering algorithms. The article presents clustering Web search results directly from a search engine as well as sets created from results for different queries. Documents were grouped using the following methods: EM and XMeans.

Keywords: Web search results clustering, documents similarity, snippets clustering

IDENTYFIKOWANIE DOKUMENTÓW PODOBNYCH W WYNIKACH WYSZUKIWANIA W SIECI WWW

Streszczenie: Przeszukiwanie sieci WWW jest niezmiernie trudnym zadaniem. Według Zamira i Etzioniego Internet to "miejsce bez struktury, niezorganizowane i zdecentralizowane". Chociaż istnieją potężne narzędzia w postaci wyszukiwarek internetowych, ich użycie staje się z czasem trudniejsze, gdyż ilość zaindeksowanych stron internetowych przekracza 1 bln [**?**] i nadal rośnie. Większość wyszukiwarek generuje wyniki posortowane według ich zgodności z treścią zapytania w postaci bardzo długich list. Takie podejście nie jest najlepszym rozwiązaniem z powodu rozmiaru list oraz zawierania w nich dokumentów nie związanych z zapytaniem. W celu zwiększenia efektywności przeszukiwania Internetu można zastosować grupowanie podobnych dokumentów z generowanej przez wyszukiwarki listy wyników. Jednym z takich narzędzi są tradycyjne algorytmy grupujące. W artykule przedstawiono wyniki

Zeszyty Naukowe Politechniki Białostockiej. Informatyka, vol. 9, pp. 61-76, 2012.

grupowania dokumentów bezpośrednio z listy zwróconej przez wyszukiwarkę oraz zbiorów dokumentów utworzonych z wyników wyszukiwania dla kilku zapytań. Wykorzystano następujące metody grupujące: EM i XMeans.

Słowa kluczowe: grupowanie wyników wyszukiwania, podobieństwo dokumentów, grupowanie snippetów

Artykuł zrealizowano w ramach pracy badawczej nr S/WI/5/08.