УДК 004.451

НАГОРЯНСЬКА В.А.**, *НІКІТІНА Л.О.*****,** к.т.н, доц., ***НІКІТІН О.О.***

АВТОМАТИЗАЦІЯ СЕМАНТИЧНОГО АНАЛІЗУ МЕДІА-ТЕКСТІВ

Електронна інформація відіграє все більшу роль у всіх сферах життя сучасного суспільства. В останні роки обсяг текстової інформації в електронному вигляді зріс настільки, що виникає загроза її знецінення, у зв'язку з труднощами пошуку необхідних відомостей серед безлічі доступних текстів. Розвиток інформаційних ресурсів Інтернет багаторазово збільшив проблему інформаційного перевантаження. Фахівці в галузі інформаційних технологій в останні роки висунули чимало ідей, присвячених скорочення часу пошуку інформації в Інтернет. Опис цих алгоритмів можна знайти в різних підручниках, періодичних виданнях та у вигляді Web-ресурсів. Нові алгоритми та їх поліпшені варіанти з'являються постійно [1], але актуальною залишається задача відбору ресурсів на основі семантичного аналізу.

Метою даної роботи є розробка та реалізація системи автоматизованого збору та тематичного аналізу новинних WEB-сторінок з компонентами накопичення знань та навчання.

Розроблена система складається з програми-агрегатора, яка збирає інформацію з різних джерел, представлену у форматі RSS-стрічок, підсистеми семантичного аналізу, бази даних з переліком тематичних категорій новинних статей, словника з вагами слів відносно тематичних категорій.

На першому етапі роботи система виконує збір та попередній аналіз новинних RSS-стрічок. Результатом попереднього аналізу є відбір URL вагомих WEB-ресурсів.

На другому етапі виконується закачка WEB-документу та проводиться його аналіз на основі методу латентно-семантичного індексування (LSI). У ході аналізу виключаються стоп-слова, будується матриця індексованих слів та виконується її сингулярне розкладання [3]. Таким чином, виконується обробка текстової інформації WEB-документу, виявляється взаємозв'язок між наявними документами і словами, які в них зустрічаються, та встановлюються ваги слів відносно наявних тематик для всіх документів. Семантичне значення документу визначається набором слів, які зазвичай йдуть разом.

Система може виконувати семантичний аналіз медіа-текстів у режимі навчання або у автоматичному режимі. У режимі навчання результати аналізу надаються експертові, який підтверджує або відкидає зміни індексів слів, пропоновані системою. У режимі автоматичного пошуку зміна індексів виконується системою.

Список літератури: 1. *Сегалович И.В*. Как работают поисковые системы //Мир Internet, - 2002. - №10. **2.** *Ландэ Д.В*. Поиск знаний в Internet. Профессиональная работа. - М.: "Вильямс", 2005. - 272 с. **3**. Латентно-семантический анализ. - http://habrahabr.ru/blogs/algorithm/110078