УДК 004.421

К.т.н., доцент Петрашенко А.В., магістрант Кобизєв С.І.

Національний технічний університет України
«Київський політехнічний інститут»

АЛГОРИТМ АВТОМАТИЧНОГО АНОТУВАННЯ ТЕКСТОВИХ ДОКУМЕНТІВ

Abstract

Andriy V. Petrashenko, assoc. prof., PhD; Sviatoslav Kobyziev, student

Automatic text summarization algorithm

This paper concerns the task of automatic text summarization – producing the most important content from a given text document to the user in a condensed form. The main goal of this research is to study an efficient method to achieve an automatic text summarization system with a high accuracy and low cost in computational times. To obtain this, we propose hybrid approach to automatic text summarization in which statistical and heuristic approaches are combined.

Вступ

Сьогодні дуже активно використовуються електронні бібліотеки. Більша частина їх еволюціонувала від традиційних бібліотек, але вони все ще не достатньо використовують переваги сучасних обчислювальних технологій. Електронні бібліотеки не мають бути тільки пасивними архівами, необхідно використовувати більш сучасні інструментальні засоби. Одним із способів розширення можливостей електронних бібліотек є створення ними і надання за вимогою користувачу метаданих. Прикладом метаданих є анотації до текстових документів. Отже, анотації є ключовим фактором для електронних бібліотек, їхнього розвитку й поширення. Але на даний час, більшість електронних бібліотек не надають послуги анотації. Анотації ще залишаються одним із критичних вузьких місць для автоматизації електронних бібліотек, тому розробка нових алгоритмів автоматичного анотування є актуальною науково-технічною задачею.

Задача автоматичного анотування текстових документів полягає у видачі користувачу найбільш важливої інформації із вхідного текстового документу.

Постановка задачі

Метою дослідження є розробка узагальненого алгоритму функціонування програмних засобів автоматичного анотування текстових документів, які виступатимуть в якості компоненту інформаційно-пошукової системи.

Опис алгоритму

Розроблений алгоритм автоматичного анотування текстових документів складається з наступних етапів :

1. Розбиття тексту документа на речення.
2. Підрахунок «ваги» кожного з отриманих речень.
3. Ваги всіх речень нормалізуються, використовуючи середнє значення ваги речення.
4. Відфільтровуються речення, вага яких не перевищує певного граничного значення.
5. Формування повної анотації із відібраних речень.
6. Формування анотації фіксованого об□ єму із повної анотації.

Ілюструє розроблений алгоритм наступна схема (рис.1).

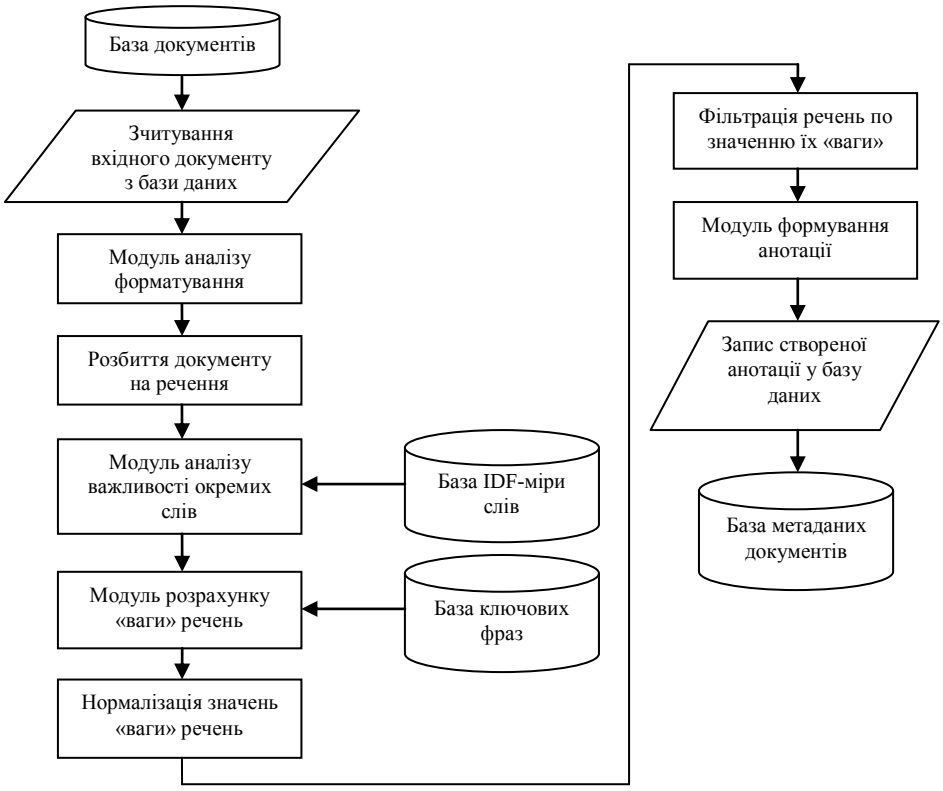

Рис.1. Узагальнена схема функціонування програмних засобів автоматичного анотування текстових документів.

Для визначення «ваги» речення пропонується використати гібридний підхід, який буде поєднувати в собі обрані ідеї із статистичного та евристичного підходів. Використання гібридного підходу дасть нам змогу отримувати більш точні результати анотування при невеликих обчислювальних витратах.

«Вага» речення W(S) визначається за формулою :

$$W(S) = k_1 * C(S) + k_2 * L(S) + k_3 * F(S) + k_4 * I(S) \ ,$$

де вага речень залежить від наступних величин :

C(S) – величина, що визначає входження ключових фраз в дане речення. Пропонується враховувати вплив входження ключових фраз, так як, на основі спостережень виявлено, що важливі та неважливі речення містять свої набори ключових фраз/слів[2]. Наприклад: для важливих речень – слово «значно», для неважливих – слова «ледве» чи «неможливо».

L(S) – величина, що визначає положення речення відносно початку документу. Існує припущення, що речення, які знаходяться на початку документу, з більшою ймовірністю будуть вадливими, ніж речення в середині документу, чи його кінці[2,3]. Незважаючи на відносну простоту цього підходу, вважається, що за допомогою даного критерію можна виділити близько 30% важливих речень тексту.

F(S) – величина, що визначає особливості форматування даного речення та вплив на його важливість форматування іншої частини документу. Визначено, що слова із заголовків та підзаголовків статті набагато частіше зустрічаються у важливих реченнях документів, ніж у неважливих[2]. Крім аналізу входжень в дане речення слів із заголовків та підзаголовків, пропонується проводити аналіз використання нестандартних стилів шрифту у даному реченні(жирний, напівжирний, курсив тощо).

I(S) – величина, що визначає ступінь важливості речення за допомогою оцінки важливості його окремих слів. Для оцінки важливості слів розраховується вага кожного слова в документі. Так як розроблені програмні засоби планується використовувати у великій інформаційній системі, ми можемо використовувати частотні методи для оцінки важливості окремих слів документу, розглядаючи кожний окремий документ як частину їх колекції. Вага окремого слова визначається як добуток частоти його входжень до даного документу(TF – term frequency) та ступеня важливості слова в контексті колекції(IDF – inverse document frequency).

$$IDF = \log \frac{|D|}{|(d_i \supset t_i)|}$$

Ступінь важливості слова представляє обернене значення частоти, з якою дане слово зустрічається в документах колекції. Врахування оберненого значення частоти зменшує вагу широковживаних слів.

$$TF = \frac{n_i}{\sum_k n_k}$$

Частота входжень слова оцінює важливість слова у рамках даного документу, визначається як співвідношення входжень даного слова до загальної кількості слів документу.

В результаті велику вагу отримують слова, що мають високу частоту входження в рамках даного документу та низьку частоту входжень в інших документах. Отримавши вагу кожного із слів документу, ми можемо визначити важливі речення по входженню в них слів з найбільшою вагою[1].

k_1 , k_2 , k_3 , k_4 – коригуючи коефіцієнти, що будуть підбиратись експериментально для балансування впливу різних факторів і отримання оптимальних результатів.

Висновки

Дослідження та розробка нових алгоритмів автоматичного анотування текстових документів є достатньо актуальною науково-технічною задачею. У ході дослідження розроблено узагальнений алгоритм автоматичного анотування із використанням гібридного підходу, який дозволяє скоротити обчислювальні витрати порівняно із стандартними підходами.

Серед загальних способів застосування анотації, окрім видачі додаткової інформації про бібліотечні ресурси, слід відзначити використання у інформаційно-пошукових системах для поліпшення функцій індексації та пошуку, а також для формування пошукового відгуку, коли в якості результатів пошуку виводиться не список документів, а список створених із текстів документів анотацій.

Література

1. *I. Mani, M. Maybury* Advances in Automatic Text summarization // The MIT Press, 1999.
2. *K. Knight, D. Marcu* Summarization beyond sentence extraction // Artificial Intelligence Випуск 139, 2002. – С. 91-107.
3. *Jing, R. McKeown* Cut and Paste text summarization // 1st Conference of the North American Chapter of the Association for Computational Linguistic, 2000. – С. 178-185.