КОМБИНИРОВАННОЕ ИСПОЛЬЗОВАНИЕ МЕТОДОВ LSA И TRM НА СЕМАНТИЧЕСКИХ УРОВНЯХ ПРОЦЕССА РЕФЕРИРОВАНИЯ

В.Е. Дручинина, О.В. Канищева

Национальный технический университет «Харьковский политехнический институт»

С ростом объема текстовой информации, доступной на WWW, стало необходимым для пользователей использовать автоматизированные инструментальные средства, чтобы находить, извлекать, фильтровать и оценивать желаемую информацию. Одним из таких средств является автоматическое реферирование текстовых документов [1, 2].

Цель задачи автоматического реферирования состоит в извлечении из документа значимых элементов (контекст, предложение, параграф), отражающих его содержание. Целью данной работы является получение автоматического реферата с использованием метода LSA и TRM для русско- и украиноязычных полнотекстовых документов.

Идея метода заключается в представлении текста в виде графа, вершинами, которыми являются предложение. Каждое предложение идентифицируется взвешенным вектором слов и вычисляется мера подобия между предложениями, определенная скалярным произведением. Если мера подобия больше заданного порога, то эти вершины соединяются. Критерий извлечения предложения в реферат определяется количеством ребер, связывающих его с другими предложениями.

Локальная характеристика предложения определяется по формуле TL*TF (Term Length*Term Frequency). Глобальная характеристика определяется методом TRM. В данной работе предлагается использовать подход LSA+TRM (Latent Semantic Analysis+TRM), который получает семантическую матрицу документа с помощью LSA. Потом, используя семантическое представление, конструирует семантический TRM.

Рассмотрим метод, основанный на использовании карты текстовых отношений (TRM – Text Relationship Map). Идея метода заключается в представлении текста в виде графа [3].

$$G = (P, E),$$

где $P = \{p_1, p_2, ..., p_k, ..., p_n\}$ – взвешенные векторы слов, соответствующие фрагментам документа. Вектор включает в себя веса составляющих его слов. Например, k-й фрагмент будет представлен вектором:

$$\{\omega_{k,1}, \omega_{k,2}, ..., \omega_{k,i}, ...\omega_{k,m}\},$$

где $\omega_{k,1}$ – вес слова, находящегося в позиции i фрагмента k; E – множество дут между узлами графа:

$$E = \{(p_k, p_b,), p_k, p_b \in V\}.$$

На рис. 1 изображен пример такой карты.

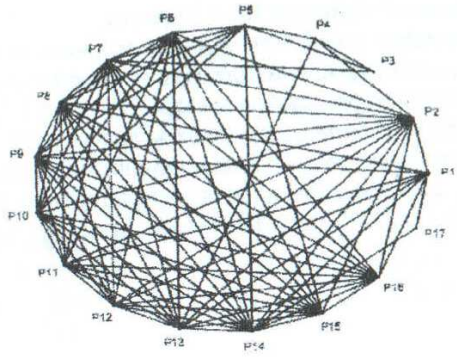

Рис. 1 – Текстовая карта предложения