Многоуровневая модель текстового сообщения, учитывающая его стилистические особенности

Multilevel model of a text message that takes into account stylistic features

Ключевые слова: информационный поиск – information retrieval; стиль – style; стилистические особенности – style features; модель – model; ключевые слова – keywords.

В статье рассматриваются вопросы информационного поиска. Описаны недостатки использования ключевых слов. Представлена многоуровневая модель текстового сообщения, учитывающая его стилистические особенности.

The paper dwells the information retrieval. Disadvantages of keywords using are described. There is presented the multilevel model of a text message that takes into account its stylistic features.

РОМАНИШИН / ROMANISHIN G.
Геннадий Валерьевич
(pochtaperep@gmail.com)
Академия ФСО России,
адъюнкт.
г. Орел

В течение последних десятилетий наблюдается постоянно ускоряющийся рост объемов текстовой информации, хранящейся в виде электронных документов. Среди средств обработки наибольшую популярность получил механизм ключевых слов, который позволяет оценить семантическую составляющую документа. Но, несмотря на простоту реализации и использования, зачастую предоставляются результаты, не удовлетворяющие пользователя.

Рассмотрим классический информационный поиск, результаты которого характеризуются релевантностью – степенью соответствия результатов поиска исходному запросу. В настоящее время много внимания отводится такой смысловой характеристике как пертинентность – соответствие полученных в результате поиска документов информационным потребностям пользователя, а не формальному соответствию документа запросу [1]. Использование только семантических данных в процессе информационного поиска, в частности механизма ключевых слов, не всегда дает требуемый результат.

Например, в статье [2] приводятся результаты поиска в сети «Internet» по запросу «крокодил», среди которых могут быть:

– энциклопедическая статья;
– биологические факты;
– новость о новой атаке в Австралии;
– запись в блоге об опыте охоты;
– произведение о жизни дикой природы и т.д.

Приведенные результаты поиска, несмотря на то, что релевантны запросу, содержат разнородную информацию, относящуюся к различным категориям. Но чаще всего пользователю необходима информация конкретной категории. Представленные результаты показаны в качестве наглядного примера. С помощью дополнительных ключевых слов (например, «научная статья» в случае поиска энциклопедической статьи) можно добиться некоторого улучшения, но в целом результаты информационного поиска существенно не изменятся. Таким образом, необходимо использовать дополнительную информацию, помимо ключевых слов, способную повысить результаты информационного поиска.

Согласно энциклопедии языка и литературы [3] текст – написанное высказывание, выходящее за рамки фразы, т.е. являющееся дискурсом и представляющее собой нечто законченное, единое и целое, наделенное внутренней структурой и организацией, соответствующей правилам какого-либо языка. Основной недостаток ключевых слов – использование большого упрощения текстового сообщения и отсутствие учета характеристик, присущего тексту в целом. Поэтому в качестве дополнительной информации необходимо использовать характеристики, описывающие текст в целом. К одной из таких характеристик относится стиль текстового сообщения.

Согласно словарю-справочнику по риторике [4] «Стиль – одна из разновидностей языка, языковая подсистема со своеобразным словарем, фразеологическими сочетаниями, оборотами, конструкциями, отличающаяся от других разновидностей в основном экспрессивно-оценочными свойствами составляющих ее элементов и обычно связанная с определенными сферами употребления речи».

В толковом словаре русского языка Дмитриева [5] приводятся такие определения:

– стилем называют индивидуальную авторскую манеру, которая ощущается читателем, зрителем в нескольких произведениях одного автора;

– стилем называют совокупность литературных приёмов, характерных для какого-либо направления, жанра, произведения.

Определение 1. Под стилем текстового сообщения в данной работе понимается языковая подсистема со своеобразным словарем, фразеологическими сочетаниями, оборотами и конструкциями, с характерной совокупностью литературных приёмов, связанная с определенными сферами употребления речи, а также определенная индивидуальная авторская манера.

Определение 2. Под стилистическими особенностями в представленной работе понимаются параметры, с помощью которых возможно описать стиль текстового сообщения.

Такими параметрами могут быть, например, согласно [6,7], средняя длина слова в буквах, средняя длина предложения (слов и букв), распределение слов по частям речи, количество сложных и простых предложений и т.д.

Исследователями было доказано, что стиль влияет на различные характеристики текста, а учет стилистических особенностей, с помощью которых производится описание стиля текстового сообщения, способен усилить поиск по ключевым словам [8].

Вследствие того, что стиль является характеристикой текста в целом и влияет на его формальные характеристики, необходимо определить уровни рассмотрения текстового сообщения. В настоящее время, несмотря на большое многообразие исследований, нет единого подхода к решению данного вопроса, однако значимую долю исследований занимают лингвистические подходы. Согласно [9] в выделении структурных уровней текста обнаруживается разноречивость, что обусловлено сложностью текста как объекта изучения и различной исходной научной позицией исследователя. Если попытаться обобщить различные подходы к выделению уровней текста, то можно представить основные подходы следующим образом (рис. 1):

Для использования на практике представленного подхода к уровням рассмотрения текста его необходимо модифицировать и выделить следующие уровни рассмотрения текстового сообщения: прагматический; семантический; синтаксический; лексический; морфологический и графематический.

Нижние уровни (графематический, морфологический, лексический и синтаксический) соответствуют одноименным уровням, выделяемым при функционально-коммуникативном подходе, за исключением фонетического, в виду того, что в письменной речи используются не фонемы, а графемы (отсюда и название первого уровня). Основными единицами обработки на данных уровнях являются единицы языка: графема, морфема, лексема, словосочетание и предложение.

Высокие уровни представлены семантическим уровнем, который включает в себя понимание и толкование слов, фраз, предложений, и

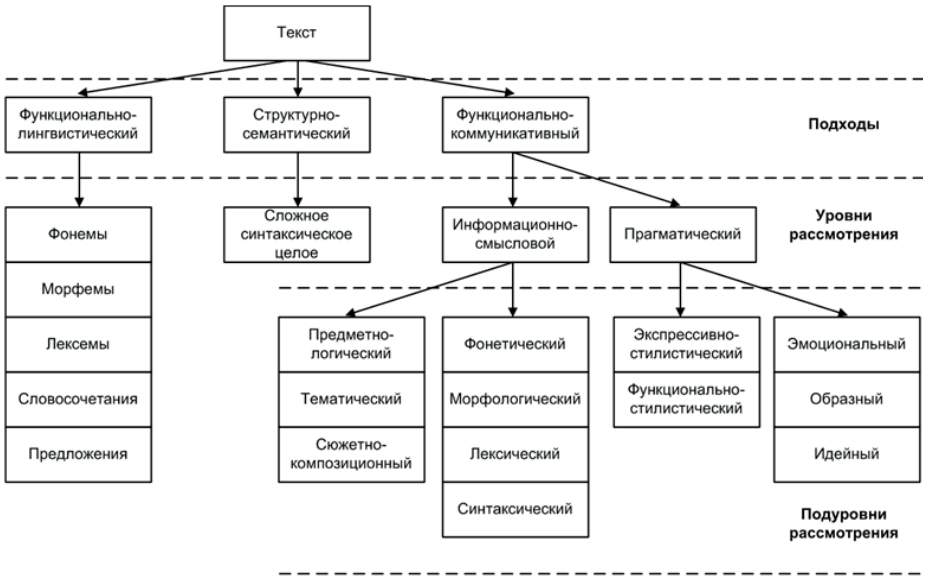

Рис. 1. Основные уровни рассмотрения текста

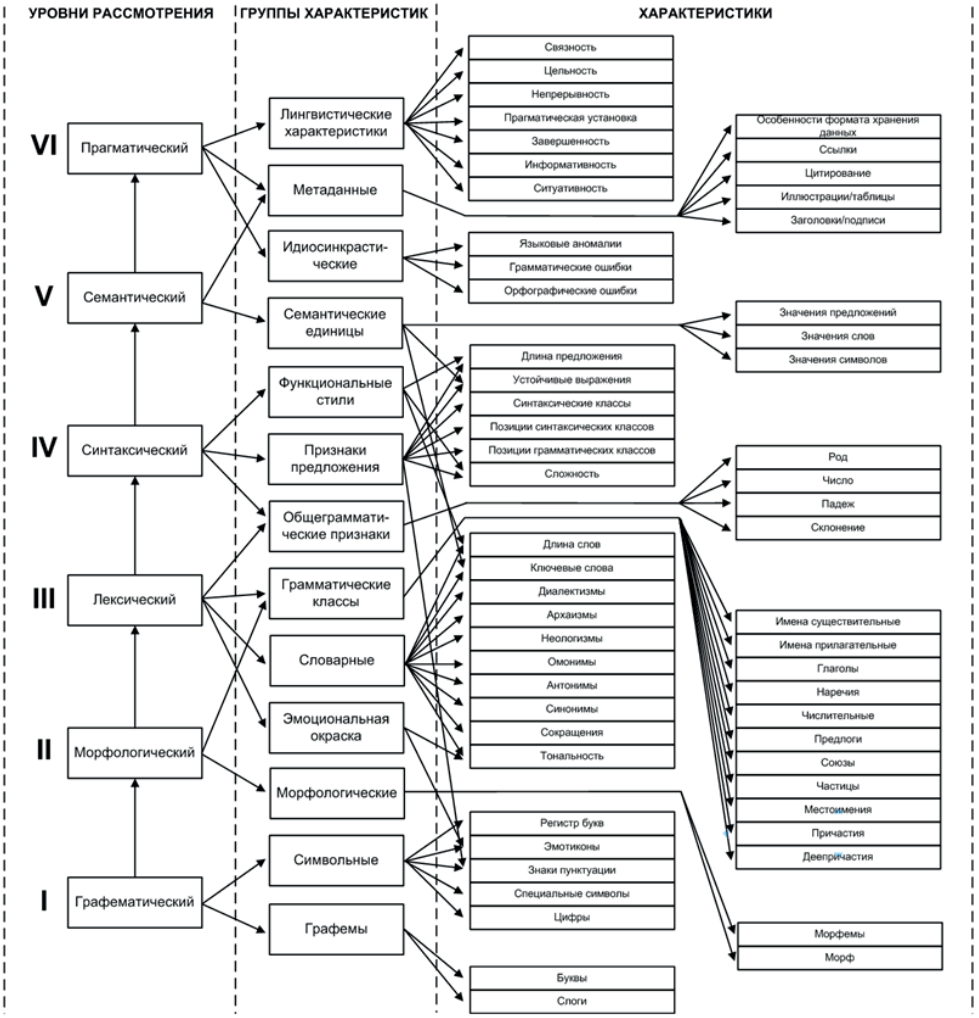

Рис. 2. Взаимосвязь уровней рассмотрения текстового сообщения и его основных характеристик

прагматическим уровнем, содержащим дополнительные знания, необходимые для обработки и понимания текста.

Отражение стилистических особенностей, характеризующих стиль текстового сообщения, оказывает влияние на его формальные характеристики. Проведя анализ работ, можно выделить следующие характеристики, а также их взаимосвязь с основными уровнями обработки текстовых сообщений (рис. 2).

Графематический уровень имеет дело с наименьшими единицами текстового сообщения – графемами и символами. Морфологический уровень касается анализа морфем и грамматических классов. На лексическом уровне анализируются слова и их различные характеристики. Синтаксический уровень содержит признаки, характеризующие предложения и правила, по которым происходит их построение. Семантический уровень включает в себя информацию о семантике основных единиц текста. К прагматическому уровню относится вся та информация, которая способствует пониманию семантики и основной идеи, заложенной в тексте.

Необходимо отметь, что при повышении уровня рассмотрения текстового сообщения усложняются характеристики текста, представляющие каждый из уровней. Представленная схема характеристик текстового сообщения достаточно полно описывает текстовое сообщение, учитывая различные уровни рассмотрения. Однако такое описание является достаточно общим и не учитывает специфику конкретной решаемой задачи. Для описания стиля текстового сообщения, отражающего характеристики текстового сообщения различных уровней представления, в рамках задачи классического информационного поиска необходимо разрабо-

тать модель текстового сообщения, учитывающую его стилистические особенности.

В рамках модели «множество слов» текст рассматривается как набор термов. Существующий механизм ключевых слов также основан на данной модели. Стоит отметить, что согласно рис. 2 ключевые слова соответствуют только лишь одному уровню рассмотрения текста – оценка семантического уровня, причем достаточно упрощенная (в зависимости от способа выделения и подсчета ключевых слов) и не всегда соответствующая истинным значениям семантических единиц. Стиль текстового сообщения охватывает все уровни его рассмотрения. Поэтому необходимо представление текстового сообщения в виде его характеристик, отражающих все уровни рассмотрения текста.

Несмотря на то, что характеристики текстового сообщения, представленные на нижних уровнях, достаточно просты в описании и вычислении, их количество достаточно велико. Для их учета предлагается использовать матрицы частот переходов, которые применяются в задачах распознавания авторства на основании авторского стиля [10]. Авторский стиль, который является составляющей стиля текстового сообщения, при использовании такого подхода определяется по признакам низких уровней рассмотрения текстового сообщения. Общий вид матрицы частот переходов представлен выражением 1:

$$M = \begin{array}{c|cccc} & N_0 & N_1 & \dots & N_k \\ \hline N_0 & m_{00} & m_{01} & \dots & m_{0k} \\ N_1 & m_{10} & m_{11} & \dots & m_{1k} \\ \dots & \dots & \dots & \dots & \dots \\ N_k & m_{k0} & m_{k1} & \dots & m_{kk} \end{array}, \quad (1)$$

где N_i – элементы, между которыми производится подсчет переходов, m_{ij} – значения переходов между N_i и N_j, $i = \overline{0..k}$, $j = \overline{0..k}$.

Согласно [10] в качестве элементов, между которыми осуществляются переходы, могут использоваться буквы, слова, предложения и т.д. Однако для оценки характеристик текста на графематическом уровне достаточно использование в качестве элементов всех символов, используемых в текстовом сообщении (буквы, цифры, специальные символы и т.д.).

Использование матриц частот переходов позволяет оценить авторскую манеру, которая является лишь частичной характеристикой стиля текстового сообщения. Для учета остальных характеристик необходимо объединить все остальные признаки в набор стилистических параметров, который задается выражением:

$$S = \{x_1, x_2, ..., x_p\}, \quad (2)$$

где x_i – составляющие набора стилистических параметров, p – количество составляющих, $i = \overline{1..p}$.

Удобство такого представления заключается в том, что при выявлении новых признаков они могут быть легко внесены в набор.

Отдельно необходимо выделить ключевые слова, на которых базируется большинство современных поисковых систем и для которых создано большое количество инструментов и способов работы с ними (использование булевой алгебры, построение индексов и т.д.) [1].

Тогда текст можно представить в следующем виде:

$$T = \{M, S, K\}, \quad (3)$$

где M – матрица частот переходов вида 1, S – набор стилистических параметров вида 2, K – ключевые слова,

Однако необходимо отметить, что не все составляющие, входящие в описание набора стилистических параметров, равнозначны. Чтобы учесть такую неравнозначность, вводятся коэффициенты значимости для каждой составляющей, которые характеризуют вклад составляющей в формирование набора стилистических параметров:

$$\mathrm{A}_S = \{\alpha_1, \alpha_2, ..., \alpha_p\}, \quad (4)$$

где α_i – коэффициенты значимости составляющих набора стилистических параметров, p – количество составляющих, $i = \overline{1..p}$.

Значения этих коэффициентов лежат в интервале [0;1]: 0 – означает минимальный вклад в формирование набора стилистических параметров, а 1 – максимальный.

Использование коэффициентов позволяет применять такой подход для различных задач, не изменяя параметры описания стиля, за счет варьирования значений коэффициентов значимости.

Применение коэффициентов значимости представляет собой механизм учета значимости признаков, который возможно применить также к матрицам частот переходов. Коэффициенты значимости для матриц частот переходов будут иметь вид:

$$\mathrm{A}_M = \begin{array}{c|cccc} & N_0 & N_1 & \dots & N_k \\ \hline N_0 & \alpha_{00} & \alpha_{01} & \dots & \alpha_{0k} \\ N_1 & \alpha_{10} & \alpha_{11} & \dots & \alpha_{1k} \\ \dots & \dots & \dots & \dots & \dots \\ N_k & \alpha_{k0} & \alpha_{k1} & \dots & \alpha_{kk} \end{array}, \quad (5)$$

где N_i – элементы, между которыми производится подсчет переходов, α_{ij} – значение коэффициента значимости для перехода между N_i и N_j, $i = \overline{0..k}$, $j = \overline{0..k}$.

Также в процессе проведения исследований было замечено, что в одном текстовом сообщении могут наблюдаться существенные скачки в стилистических характеристиках. Так, в газетной статье, описывающей политическое событие, используется газетный стиль. Но, например, в эпиграфе приводится стихотворение знаменитого писателя, в той или иной форме характеризующее описываемое в статье событие. Это уже использование художественного стиля. Далее могут встречаться выдержки из нормативно-правовых документов, что соответствует уже официально-деловому стилю и т.д. Для учета таких неоднородностей стилистических особенностей предлагается использовать разбиение текста на зоны и учитывать не только стилистические параметры всего текста, но и стилистические параметры текста в каждой зоне, по аналогии с «локальным множеством слов», представленным в работе [11].

Локальный набор стилистических параметров имеет вид:

$$S_L = \{S_{L_1}, S_{L_2}, ..., S_{L_z}\}, \quad (6)$$

где S_{L_i} – набора стилистических параметров для i-й зоны, z – количество зон, $i = \overline{1..z}$.

Аналогично задается локальные матрицы частот переходов:

$$M_L = \{M_{L_1}, M_{L_2}, ..., M_{L_z}\}, \quad (7)$$

где M_{L_i} – матрица частот переходов для i-й зоны, z – количество зон, $i = \overline{1..z}$.

Тогда с учетом выражений 1–7, модель текстового сообщения, учитывающая его стилистические особенности, будет иметь следующий вид:

$$T = \{M, M_L, S, S_L, K, \mathrm{A}_M, \mathrm{A}_S\}, \quad (8)$$

где M – матрица частот переходов для текстового сообщения в целом;

M_L – множество локальных матрица частот переходов;

S – набор стилистических параметров для текстового сообщения в целом;

S_L – локальный набор стилистических параметров;

A_M – коэффициенты значимости матриц частот переходов;

A_S – коэффициенты значимости набора стилистических параметров.

Таким образом, представленная многоуровневая модель текстового сообщения, учитывающая его стилистические особенности, позволяет учесть характеристики текстового сообщения на различных уровнях представления, устраняет основной недостаток ключевых слов и способствует улучшению процесса ведения информационного поиска.

Литература

1. *Ландэ, Д.В.* Интернетика: Навигация в сложных системах: модели и алгоритмы. / Д.В. Ландэ, А.А. Снарский, И.В. Безсуднов – М.: Книжный дом «ЛИБРОКОМ», 2009. – 264 с.

2. *Petrenz, P.* Cross-Lingual Genre Classification / P. Petrenz // Proceedings of the Student Research Workshop at the 13th Conference of the European Chapter of the Association for Computational Linguistics – 2012. – P. 11–21.

3. Литература и язык. Современная иллюстрированная энциклопедия. / под ред. А.П. Горкина – М.: Росмэн, 2006. – 584 с.

4. *Жеребило, Т.В.* Риторика: Словарь-справочник. / Т.В.Жеребило – Назрань: Пилигрим, 2011. – 56 с.

5. *Дмитриев, Д.В.* Толковый словарь Русского Языка / Д.В. Дмитриев – М.: Астрель: АСТ, 2003. – 1578 с.

6. *Браславский, П.И.* Методы повышения эффективности поиска научной информации (на материалу Internet): дис. … канд. техн. наук: 05.13.16 // Павел Исаакович Браславский. Уральский государственный технический университет – Екантеринбург, 2000. – 154 с.

7. *Karlgren, J.* Stylistic Experiments for Information Retrieval: A Dissertation submitted for the Degree of Doctor of Philosophy in Computational Linguistics // Jussi Karlgren – Stockholm, Sweden, 2000. – 147 p.

8. *Браславский, П.И.* Стиль как дополнительный параметр поиска информации в Internet / П.И. Браславский. // Русская компьютерная и квантитативная лингвистика. – 2000. – С. 396.

9. *Бабенко, Л.Г.* Лингвистический анализ художественного текста / Л.Г.Бабенко, Ю.В. Казарин – М.: Флинта: Наука, 2005. – 496 с.

10. *Поддубный, В.В.* Сравнительный анализ эффективности алгоритмов распознавания авторства текстов по частотам переходов / В.В. Поддубный, О.Г.Шевелев, А.А. Фатыхов // Вестн. Том. гос. ун-та. – 2006. – № 290. – С. 232–234.

11. Local Word Bag Model for Text Categorization / Wen Pu [и др.] // Data Mining, ICDM 2007, Seventh IEEE International Conference. – 2007. – P. 625–630.