УДК 004.912
DOI: 10.22213/2410-9304-2019-2-58-64

ОБ ОДНОМ ПОДХОДЕ К ПОСТРОЕНИЮ ИНФОРМАЦИОННОЙ СИСТЕМЫ ОБРАБОТКИ ТЕКСТОВОЙ ИНФОРМАЦИИ НА ОСНОВЕ СМЫСЛОВЫХ ГРУПП

С. В. Моченов, кандидат технических наук, профессор, ИжГТУ имени М. Т. Калашникова, Ижевск, Россия
Р. Р. Ахметгалеев, аспирант, ИжГТУ имени М. Т. Калашникова, Ижевск, Россия

В статье рассматривается подход к анализу текста, основанный на построении и использовании баз данных частей речи и других членов предложения. Соответствующие базы данных для русскоязычных текстов формируются на основе экспертных оценок, получаемых в процессе анализа текстовых массивов с предложениями различной сложности. Актуальность работы связана с проблемой автоматизации поиска и выделения полезной для пользователя информации, необходимой для решения конкретных задач. В процессе анализа формируются различные массивы индексов. Осуществляется: выделение различных сочетаний слов предложения, сравнение их с допустимыми комбинациями из базы данных (формирование смысловых групп), структуризация предложений и формирование иерархической системы смысловых групп. На примерах показаны развернутые результаты работы программного комплекса. При анализе основных частей предложения (темы и ремы) используется одинаковый набор функциональных модулей. Представленные результаты показали принципиальную возможность создания подобной информационной системы анализа текстовой информации на основе изложенного подхода. Разработанный программный комплекс при выделении СГ анализирует комбинации слов, а не отдельные предлоги, союзы и другие вспомогательные элементы предложений. За счет разделения на СГ с использованием экспертных баз данных обеспечивается более полное сохранение смысловой составляющей текста. В дальнейшем предполагается расширение сферы применения программного комплекса для выделения полезной для пользователя информации, сокращения ее объема, уменьшения времени, затрачиваемого на поиск.

Ключевые слова: информационная система, обработка текстовой информации, смысловые группы, сокращение текста, смысловая составляющая, выделение полезной информации.

Введение

Автоматизация обработки текстовой информации особенно актуальна в настоящее время, поскольку информационные потоки по различным направлениям человеческой деятельности растут очень быстро. Накопленный опыт отражается в многочисленных текстах, хранящихся в библиотеках и в глобальной сети Интернет. Пользователь при решении своих научных или практических задач сталкивается с проблемой поиска и выявления нужной ему информации [1–4].

С одной стороны, автор научной статьи или публикации, используя свой опыт, формирует логическую и содержательную структуру в виде последовательности предложений, абзацев, разделов в тексте публикации. Смысловая составляющая текста последовательно складывается и обобщается на основе смысловых составляющих перечисленных компонентов текста.

С другой стороны, исследователь (читатель, пользователь), ориентируясь на свои интересы, пытается найти в той или иной публикации полезные для него сведения, что требует значительных затрат времени и интеллектуальных затрат.

Разработка интеллектуальных систем обработки текстовой информации позволит существенно сократить эти затраты. Представляется актуальным использование при разработке подобных систем концепции смысловых групп (СГ).

Описание методов исследования

В работе [5] для анализа текста введено понятие вектора цели. Это понятие применимо как к отдельным предложениям, так и к более крупным текстовым структурам. Каждое предложение в тексте имеет свое предназначение и служит некоторым промежуточным результатом, отражающим часть смысловой нагрузки, которую автор передает в своей публикации. Предложена модель вектора цели, основанная на основных ролевых функциях предложения: связующей, структурной, семантической.

Вектор цели предложения выделяется на семантическом уровне. Векторы цели абзацев, отдельных глав или всего текста формируются по результатам выделения векторов целей предложений. В качестве координат вектора цели могут выступать отдельные ключевые слова, смысловые группы предложений, отдельные предложения или абзацы.

В ряде работ [6–10] рассматривается информационный подход к анализу текста, основанный на разбиении предложения (а в дальнейшем и всего текста) на составные части: тему и рему. Тема связана с ответом на вопрос «О чем говорится?», а рема – с ответом на вопрос «Что говорится?».

© Моченов С. В., Ахметгалеев Р. Р., 2019

В работе [11] в качестве условного разделителя на тему и рему в предложении используется первый встреченный глагол. При этом левая часть от глагола считается темой, а правая – ремой. При анализе отдельных предложений разбиение на тему и рему можно рассматривать как первый этап структуризации предложения.

Следует отметить, что при подобном разбиении могут возникнуть ситуации, связанные либо вообще с отсутствием в предложении глагола, либо с расположением его по концам предложения. Кроме того, в сложных предложениях может встретиться несколько глаголов. Разрешение этих проблем осуществляется алгоритмически.

Введение понятия смысловых групп разбивает исходное предложение на законченные смысловые комбинации слов. В роли элементов разделителей на смысловые группы выступают предлоги, союзы, знаки препинания, вспомогательные слова, взаимное расположение отдельных частей речи с предлогами и союзами. Разбиение на тему и рему в сочетании с выделением СГ показало хорошие результаты, приведенные в работе [12].

Известно, что речевое общение возникло значительно раньше письменного, текстового. Отметим, например, что обучение детей языку, речи основано не на морфологии, синтаксисе или семантике языка. О них ребенок не имеет никакого представления. Обучение разговорному языку происходит на примерах, на общении с родителями или учителем. Отдельные слова, сочетания слов, образы, связанные с ними, – это те ключевые моменты, которые лежат в основе речевых коммуникаций. При этом в речевой коммуникации между людьми дополнительно может передаваться и эмоциональная составляющая, которая в текстовых документах выражается либо подбором соответствующих СГ, либо через знаки препинания.

Перечисленные факты натолкнули авторов данной статьи на идею использования при анализе текстовой информации наборов допустимых в речевом общении комбинаций частей речи, предлогов, союзов, частиц, знаков препинания и других элементов.

На основе предварительной экспертной оценки были выявлены наиболее часто встречающиеся в русскоязычных текстах комбинации частей речи и других членов предложения. Эти комбинации предлагается использовать для выделения смысловых групп (СГ).

В качестве критериев для выявления допустимых комбинаций использованы критерии, аналогичные критериям формирования человеком речевой последовательности слов в предложении: кратковременная задержка при переходе от одной законченной смысловой группы слов к другой (пауза), интонационное подчеркивание отдельных фраз (ударение), снижение уровня интонационного подчеркивания к концу смысловой группы слов.

Было выяснено, что применение подобных критериев позволяет провести разбиение предложения на смысловые группы более точно. При этом СГ, полученные с использованием данных критериев, не всегда совпадают с СГ, получаемыми на основе предлогов, союзов и других вспомогательных слов, т. е. на основе принятых правил синтаксиса и грамматики русского языка. Подобный анализ, на основе разбиения на СГ, позволяет более полно передать смысловую информацию, заключенную в отдельных предложениях текста, сформированных автором.

Разбиение на СГ одновременно позволяет выявить уровни значимости СГ в отдельном предложении, их вложенность и связность друг с другом. Предполагается, что в дальнейшем СГ, полученные для разных участков текста, позволят связать в единую структуру смысловую составляющую, передаваемую автором в тексте.

Поскольку каждой части речи соответствует огромное множество слов, то использование допустимых комбинаций для формирования СГ позволяет унифицировать выделение СГ при анализе предложений различной сложности. При этом существенно уменьшается сложность алгоритмов анализа, поскольку отпадает необходимость проверки типа предложной и падежной сочетаемости слов.

Алгоритмически процедура анализа включает в себя следующие этапы:

1. Отнесение каждого слова к определенной части речи или члену предложения (индексация слов по частям речи).

2. Разбиение предложения на тему и рему.

3. Использование исходных баз данных допустимых комбинаций частей речи (и других вспомогательных слов) для выделения СГ. Формирование массива индексов для выделения СГ по теме и реме.

4. Формирование СГ по теме и реме.

Анализ предложения по каждой базе осуществляется на основе реализации унифицированных функций (модулей) с разным набором входных параметров. Предусмотрена возможность расширения состава исходных баз и воз-

можность настройки по интересам пользователя. Подобная организация предполагает дальнейшую реализацию на основе нейронных сетей с целью построения интеллектуальных систем, связанных с анализом больших информационных массивов текстов.

Полученные в ходе экспериментального исследования данные и их интерпретация

Ниже приведена таблица исходных баз допустимых комбинаций частей речи и других членов предложения, полученная на основе экспертных оценок.

Таблица исходных баз допустимых комбинаций частей речи для выделения смысловых групп

baza12 = '12'
baza11 = '11 11 10 11 11 11 11 9 11 9 8 11 11 9 8 11 0 8 8 8'
baza10 = '10 10 8 10 10 10 0 10 8 8 10 0 8 10 9 8 10 0 4 10 3 9 10 8 1 8 10 8 8 8 10 0 9 8 10 0 8 8 10 10 0 8 10 0 8 9 10 0 8 6 8 10 8 8 1 8 10 0 8 8 11 10 0 8 8 8 10 0 8 9 8'
baza9 = '9 8 9 9 9 9 8 9 10 8 9 8 8 9 0 4 8 9 11 10 10 9 8 1 8 9 0 9 8 9 1 9 8 9 9 1 98'
baza8 = '8 8 8 8 8 8 9 8 8 1 8 8 1 4 8 8 1 8 8 9 9 8 8 9 8 8 8 8 8 1 8 8 8 10 0 9 8'
baza7 = '7'
baza6 = '6 '
baza5 = '5 9 8 8'
baza4 = '4'
baza3 = '3 10 3 9 3 10 10 3 0 8'
baza2 = '2 9 8'
baza1 = '1 1 8 1 9 8 1 0 8 1 10 0 8 1 0 9 8 1 9 8 8 1 11 11 10 1 8 8 8 1 10 8 1 8 1 8 8 1 8 8'
baza0 = '0 4 0 8 0 10 0 8 8 0 8 9 0 9 8 0 9 10 0 11 8 0 8 8 8 0 4 9 8 0 8 6 8 0 8 9 8 0 9 8 8 0 9 9 8 0 8 8 8 8 0 9 8 8 8 0 9 8 9 8 0 8 8 9 8 0 8 0 9 8 0 8 8 8 8 8'

В приведенной таблице цифры от 0 до 12 определяют соответствующую часть речи, член предложения или знак препинания. Например, цифре 0 соответствует предлог, цифре 8 – существительное, цифре 10 – глагол, цифре 7 – запятая или точка. Наборы комбинаций в каждой базе соответствуют числу элементов в СГ и определяют состав СГ. Особенностью использования приведенной базы является то, что это фактически правила построения смысловых групп предложений, база знаний. При этом относительно малый объем базы накрывает при анализе значительно больший объем возможных вариантов построения предложений. По мере накопления статистических данных допускается возможность расширения базы.

На рис. 1–4 приведены примеры разбиения исходных предложений на тему и рему, а также на смысловые группы. Используются формируемые в процессе анализа индексы частей речи и индексы для определения СГ.

Программный комплекс как информационная система реализован с использованием среды разработки JetBrains PyCharm Community Edition 2017.3 x64, Python36-32.

Пример № 1

['поэтому', ',', 'в', 'этом', 'случае', ',', 'требуется', 'выявление', 'выделение', 'и', 'перечисление', 'основных', 'тем', 'и', 'объектов', ',', 'которым', 'посвящается', 'документ', '.'] - исходное предложение S13_1

['требуется', 'выявление', 'выделение', 'и', 'перечисление', 'основных', 'тем', 'и', 'объектов', ',', 'которым', 'посвящается', 'документ', '.'] - рема, S_rema1 исходного предложения

[10, 8, 8, 1, 8, 9, 8, 1, 8, 7, 9, 10, 8, 7] - индексы частей речи по реме

['поэтому', ',', 'в', 'этом', 'случае', ','] - тема, S_tema1 исходного предложения

[11, 7, 0, 9, 8, 7] - индексы частей речи по теме

ФОРМИРОВАНИЕ СМЫСЛОВЫХ ГРУПП РЕМЫ

[0, 5, 9, 10, 13, 14] ind1 - индексы для определения СГ ремы

['требуется', 'выявление', 'выделение', 'и', 'перечисление'] Srema0

['основных', 'тем', 'и', 'объектов'] Srema1

[','] Srema2

['которым', 'посвящается', 'документ'] Srema3

['.'] Srema4
КОНЕЦ 5
ФОРМИРОВАНИЕ СМЫСЛОВЫХ ГРУПП ТЕМЫ
[0, 1, 2, 5, 6] ind1t - индексы для определения СГ темы
['поэтому'] Stema0
[','] Stema1
['в', 'этом', 'случае'] Stema2
[','] Stema3
КОНЕЦ 4
Process finished with exit code 0

Рис. 1. Предложение 1. Смысловые группы темы и ремы

Пример № 2

['поэтому', ',', 'требуется', 'в', 'этом', 'случае', 'открыть', 'документ', '.'] - исходное предложение S13_1
['требуется', 'в', 'этом', 'случае', 'открыть', 'документ', '.'] - рема, S_rema1 исходного предложения
[10, 0, 9, 8, 10, 8, 7] - индексы частей речи по реме
['поэтому', ','] - тема, S_tema1 исходного предложения
[11, 7] - индексы частей речи по теме

ФОРМИРОВАНИЕ СМЫСЛОВЫХ ГРУПП РЕМЫ
[0, 4, 6, 7] ind1 - индексы для определения СГ ремы ['требуется', 'в', 'этом', 'случае'] Srema0
['открыть', 'документ'] Srema1
['.'] Srema2
КОНЕЦ 3
ФОРМИРОВАНИЕ СМЫСЛОВЫХ ГРУПП ТЕМЫ
[0, 1, 2] ind1t - индексы для определения СГ темы
['поэтому'] Stema0
[','] Stema1
КОНЕЦ 2
Process finished with exit code 0

Рис. 2. Предложение №2. Смысловые группы темы и ремы

Пример № 3

['перед', 'началом', 'ремонта', 'дом', 'не', 'был', 'осмотрен', 'совместно', 'с', 'представителями', 'ооо', 'союз', 'и', 'председателем', 'совета', 'дома', 'на', 'предмет', 'составления', 'дефектной', 'ведомости', 'и', 'определения', 'объема', 'и', 'стоимости', 'работ', '.'] - исходное предложение S13_1
['не', 'был', 'осмотрен', 'совместно', 'с', 'представителями', 'ооо', 'союз', 'и', 'председателем', 'совета', 'дома', 'на', 'предмет', 'составления', 'дефектной', 'ведомости', 'и', 'определения', 'объема', 'и', 'стоимости', 'работ', '.'] - рема, S_rema1 исходного предложения
[3, 10, 10, 11, 0, 8, 8, 8, 1, 8, 8, 8, 0, 8, 8, 9, 8, 1, 8, 8, 1, 8, 8, 7] - индексы частей речи по реме
['перед', 'началом', 'ремонта', 'дом'] - тема, S_tema1 исходного предложения
[0, 8, 8, 8] - индексы частей речи по теме

ФОРМИРОВАНИЕ СМЫСЛОВЫХ ГРУПП РЕМЫ
[0, 3, 8, 12, 17, 23, 24] ind1 - индексы для определения СГ ремы
['не', 'был', 'осмотрен'] Srema0
['совместно', 'с', 'представителями', 'ооо', 'союз'] Srema1
['и', 'председателем', 'совета', 'дома'] Srema2
['на', 'предмет', 'составления', 'дефектной', 'ведомости'] Srema3
['и', 'определения', 'объема', 'и', 'стоимости', 'работ'] Srema4
['.'] Srema5
КОНЕЦ 6
ФОРМИРОВАНИЕ СМЫСЛОВЫХ ГРУПП ТЕМЫ
[0, 4] ind1t - индексы для определения СГ темы

['перед', 'началом', 'ремонта', 'дом'] Stema0
КОНЕЦ 1
Process finished with exit code 0

Рис. 3. Предложение № 3. Смысловые группы темы и ремы

Пример № 4

['наши', 'требования', 'по', 'своевременному', 'предоставлению', 'технической', 'и', 'финансовой', 'документации', 'руководство', 'ооо', 'союз', 'игнорировало', ',', 'что', 'привело', 'к', 'нарушениям', 'и', 'несоблюдению', 'необходимых', 'при', 'капитальном', 'ремонте', 'норм', 'и', 'правил', '.'] - исходное предложение S13_1

['игнорировало', ',', 'что', 'привело', 'к', 'нарушениям', 'и', 'несоблюдению', 'необходимых', 'при', 'капитальном', 'ремонте', 'норм', 'и', 'правил', '.'] - рема, S_rema1 исходного предложения

[10, 7, 1, 10, 0, 8, 1, 8, 9, 0, 9, 8, 8, 1, 8, 7] - индексы частей речи по реме

['наши', 'требования', 'по', 'своевременному', 'предоставлению', 'технической', 'и', 'финансовой', 'документации', 'руководство', 'ооо', 'союз'] - тема, S_tema1 исходного предложения

[9, 8, 0, 9, 8, 9, 1, 9, 8, 8, 8, 8] - индексы частей речи по теме

ФОРМИРОВАНИЕ СМЫСЛОВЫХ ГРУПП РЕМЫ
[0, 1, 2, 6, 8, 12, 15, 16] ind1 - индексы для определения СГ ремы
['игнорировало'] Srema0
[','] Srema1
['что', 'привело', 'к', 'нарушениям'] Srema2
['и', 'несоблюдению'] Srema3
['необходимых', 'при', 'капитальном', 'ремонте'] Srema4
['норм', 'и', 'правил'] Srema5
['.'] Srema6
КОНЕЦ 7
ФОРМИРОВАНИЕ СМЫСЛОВЫХ ГРУПП ТЕМЫ
[0, 2, 5, 9, 12] ind1t - индексы для определения СГ темы
['наши', 'требования'] Stema0
['по', 'своевременному', 'предоставлению'] Stema1
['технической', 'и', 'финансовой', 'документации'] Stema2
['руководство', 'ооо', 'союз'] Stema3
КОНЕЦ
Process finished with exit code 0

Рис. 4. Предложение № 4. Смысловые группы темы и ремы

Выводы

Представленные результаты показали принципиальную возможность создания информационной системы анализа текстовой информации на основе изложенного подхода. Разработанный программный комплекс при выделении СГ анализирует комбинации элементов предложения, а не отдельные предлоги, союзы и другие вспомогательные элементы предложений.

Пользователю предоставляется результат структуризации каждого предложения и полный состав его смысловых групп. Смысловые группы выстраиваются по иерархическому принципу в соответствии с логикой построения предложения.

За счет разделения на СГ с использованием экспертных баз данных обеспечивается более полное сохранение смысловой составляющей текста.

В дальнейшем предполагается расширение сферы применения программного комплекса для выделения полезной для пользователя информации, сокращения ее объема, уменьшения времени, затрачиваемого на поиск.

Подобная организация системы предполагает возможность ее реализации на основе нейронных сетей с целью построения интеллектуальных систем, связанных с анализом больших информационных массивов текстов.

Библиографические ссылки

1. *Алексеев А. А.* Тематический анализ новостного кластера как основа для автоматического аннотирования // Программная инженерия. 2014. № 3. С. 41–48.

2. *Артюхин В. В., Чяснавичюс Ю. К.* Планирование аналитического исследования при помощи методов анализа качественных данных // Прикладная информатика. 2014. № 2. С 23–48.

3. *Герте Н. А., Курушин Д. С., Нестерова Н. М.* Моделирование понимания текста как основа автоматизированного реферирования // Материалы VII Международной научной конференции «Индустрия перевода» (Россия, Пермь, 1–3 июня 2015 г.). С. 81–84.

4. *Бледнов А. М., Моченов С. В., Луговских Ю. А.* Об одном методе статистической фильтрации текстовой информации // Современные информационные технологии и письменное наследие: от древних рукописей к электронным текстам : материалы Междунар. науч. конф. (Россия, Ижевск, 13–17 июля 2006 г.). С. 126–130.

5. *Бледнов А. М., Моченов С. В., Луговских Ю. А.* Векторная модель представления текстовой информации // Современные информационные технологии и письменное наследие от древних рукописей к электронным текстам : материалы Междунар. науч. конф. (Ижевск, 13–17 июля 2006 г). С. 136–145.

6. *Rankel P., Conroy J., Dang H., Nenkova A.* A Decade of Automatic Content Evaluation of News Summaries: Reassessing the State of the Art // Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. 2013. Pp. 131-136.

7. *Захарова И. С., Филиппова Л. Я.* Основы информационно-аналитической деятельности: учебное пособие. Киев : Центр учебной литературы, 2013. 336 с.

8. *Курушин Д. С., Нестерова Н. М., Овчинникова И. Г.* О возможном подходе к созданию системы автоматического реферирования // Вопросы психолингвистики. М., 2014. № 2 (20). С. 123–128.

9. [Abstracts - The Writing Center] [Электронный ресурс]. URL: http://writingcenter.unc.edu/handouts/abstracts/ (дата обращения: 02.04.2018)

10. *Осипов Г. С., Шелманов А. О.* Метод повышения качества синтаксического анализа на основе взаимодействия синтаксических и семантических правил // Труды шестой Международной конференции «Системный анализ и информационные технологии». Т. 1. 2015. С. 229–240.

11. *Моченов С. В., Втюрин М. В., Ахметгалеев Р. Р.* К вопросу о построении информационной системы обработки текстовой информации на основе смысловых групп // Вестник ИжГТУ имени М. Т. Калашникова. 2018. Т. 21. № 3. С. 166–170.

12. Там же.

References

1. Alekseev A.A. [Thematic representation of a news cluster as a basis for automatic annotation]. *Programmnaja inzhenerija*, 2014, pp. 41-48 (in Russ.).

2. Artjuhin V.V., Chjasnavichjus Ju.K. [Planning an analytical study using qualitative data analysis methods]. *Prikladnaja informatika*, 2014, pp. 23–48. (in Russ.).

3. Gerte N.A., Kurushin D.S., Nesterova N.M. *Modelirovanie ponimaniya teksta kak osnova avtomatizirovannogo referirovaniya* [Modeling the understanding of text as the basis for automated abstracting]. *Materialy VII Mezhdunarodnoi nauchnoi konferentsii «Industriya perevoda» (Rossiya, Perm', 1–3 iyunya 2015 g.).* [Proceedings of the VII International Scientific Conference "Translation Industry" (Russia, Perm, June 1–3, 2015)], 2015, pp. 81-84 (in Russ.).

4. Blednov A.M., Mochenov S.V., Lugovskikh Yu.A. *Ob odnom metode statisticheskoi fil'tratsii tekstovoi informatsii* [About one method of statistical filtering of textual information]. *Sovremennye informatsionnye tekhnologii i pis'mennoe nasledie: ot drevnikh rukopisei k elektronnym tekstam* [Proc. Modern information technologies and written heritage: from ancient manuscripts to electronic texts: materials of the Intern. scientific conf. (Russia, Izhevsk, July 13-17, 2006).], 2006, pp. 126-130 (in Russ.).

5. Blednov A.M., Mochenov S.V., Lugovskikh Yu.A. (2006) *Vektornaja model' predstavlenija tekstovoj informacii* [Vector model of text information representation]. *Sovremennye informatsionnye tekhnologii i pis'mennoe nasledie: ot drevnikh rukopisei k elektronnym tekstam* [Proc. Modern information technologies and written heritage: from ancient manuscripts to electronic texts: materials of the Intern. scientific conf. (Russia, Izhevsk, July 13-17, 2006).], 2006, pp. 136-145 (in Russ.).

6. Rankel P., Conroy J., Dang H., Nenkova A. A Decade of Automatic Content Evaluation of News Summaries: Reassessing the State of the Art // Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. 2013. pp. 131-136.

7. Zakharova I.S., Filippova L.Ya. *Osnovy informatsionno-analiticheskoi deyatel'nosti: uchebnoe posobie* [Fundamentals of Information and Analytical Activities]. Kiev, *Tsentr uchebnoi literatury*, 2013, p. 336 (in Russ.).

8. Kurushin D.S., Nesterova N.M., Ovchinnikova I.G. [About possible approach to creating a system of automatic summarization]. *Voprosy psikholingvistiki*, 2014, pp. 123-128 (in Russ.).

9. [Abstracts - The Writing Center] available at http://writingcenter.unc.edu/handouts/abstracts/ (accessed: April 4, 2018).

10. Osipov G.S., Shelmanov A.O. *Metod povysheniya kachestva sintaksicheskogo analiza na osnove vzaimodeistviya sintaksicheskikh i semanticheskikh pravil* [Method for improving the quality of syntactic analysis based on the interaction of syntactic and semantic rules]. *Sistemnyi analiz i informatsionnye tekhnologii* [Proceedings of the sixth International Conference "System Analysis and Information Technologies"], 2015, vol. 1, pp. 229-240 (in Russ.).

11. Mochenov S.V., Vtyurin M.V., Ahmetgaleev R.R. [To the Question of Developing an Information System for Processing Textual Information on the Basis of Semantic Groups]. *Vestnik IzhGTU imeni M. T.*

Kalashnikova. 2018. Vol. 21. No. 3. Pp. 166-170 (in Russ.).
12. Ibid.

* * *

On One Approach to Building An Information System for Processing Text Information Based on Semantic Groups

S. V. Mochenov, PhD in Engineering, Professor, Kalashnikov ISTU
R. R. Akhmetgaleev, Post-graduate, Kalashnikov ISTU

The paper considers an approach to text analysis based on the construction and use of databases of parts of speech and other members of a sentence. Corresponding databases for Russian texts are formed on the basis of expert assessments obtained in the process of analyzing text arrays with suggestions of varying complexity. The relevance of the work is related to the problem of automating the search and highlighting useful information for the user that is needed to solve specific problems. In the process of analysis, various index arrays are formed. Selection of various combinations of words of the sentence, comparing them with valid combinations of the database (the formation of semantic groups), structuring sentences, and formation of a hierarchical system of semantic groups are carried out. The examples show the detailed results of the software package. When analyzing the main parts of the sentence (themes and rhemes), the same set of functional modules is used. The presented results showed the fundamental possibility of creating such an information system for analyzing textual information based on the outlined approach. The developed software package when selecting the SG analyzes word combinations, rather than individual prepositions, conjunctions and other auxiliary elements of sentences. Due to the division on the SG using expert databases, a more complete preservation of the semantic component of the text is provided. In the future, it is intended to expand the scope of application of the software system to highlight useful information for the user, reduce its volume, reduce the time spent on search.

Keywords: information system, text processing, semantic groups, text abbreviation, semantic component, selection of useful information.

Получено: 06.05.19