© Гольдштейн С.Л.

УДК 004.81+614

СИСТЕМНОЕ ПРОТОТИПИРОВАНИЕ ФИЛЬТРА ТЕКСТОВ, РЕЛЕВАНТНО-ПЕРТИНЕНТНЫХ ДЕЯТЕЛЬНОСТИ СПЕЦИАЛИСТА МЕДИЦИНСКОГО УЧРЕЖДЕНИЯ

Гольдштейн С.Л.[1], Грицюк Е.М.[2], Кудрявцев А.Г.[1], Ходенева М.А.[1]
1 – УрФУ, 2 – МКМЦ «Бонум»

Резюме. Представлены материалы системного прототипирования фильтра текстов по «образцу» (релевантно-пертинентных деятельности специалиста медицинского учреждения). Проведен литературно-аналитический обзор по системе фильтрации, ее подсистемам, блокам, отдельным подблокам, узлам и модулям. Составлены пакеты аналогов, научных и корпоративных прототипов.
Ключевые слова: служебный текст, фильтр текстов, структура фильтра текстов, аналоги, прототипы.

SYSTEM PROTOTYPING OF A TEXT FILTER, RELEVANT-PERPINENT ACTIVITIES OF A MEDICAL INSTITUTIONAL SPECIALIST

Goldshtein S.L.[1], Gritsyuk E.M.[2], Kudryavtsev A.G.[1], Hodeneva M.A.[1]
1 – UrFU, 2 – MCMC "BONUM"

Summary. The materials of system prototyping of the filter of texts on "sample" (relevant-pertinent activity of the specialist of medical institution) are presented. A literary and analytical review of the filtration system, its subsystems, blocks, individual sub-blocks, nodes and modules was carried out. Packages of analogs, scientific and corporate prototypes have been compiled.
Keywords: service text, text filter, text filter structure, analogues, prototypes.

Введение

Ранее [1] нами сформулирована и рассмотрена задача компьютерной поддержки анализа и разрешения служебной ситуации прерывания плановой работы специалиста спонтанным текстовым сообщением, которое может отличаться по смыслу от интересов специалиста, и получены первые результаты: графическая модель основных событий за 8-ми-часовой рабочий день специалиста, столкнувшегося с такой ситуацией; фрагмент иерархии понятий по термину «Конфликтная ситуация служебной деятельности»; оценки затрат времени на работу с текстом прерывания и его значимостью в вариантах без и с компьютерным помощником; гипотезы о возможной реализации компьютерного помощника, в основе которого – фильтр текстов, точнее, система фильтрации текстов. Затем [2] представлены структура этой системы по «образцу» и ее математическая модель; блоки извлечения информации из специалиста и из текстов и их сопоставления, а также - формула сходства

сопоставляемых векторов информации. На этой основе возможна постановка дальнейших исследований.

В данной статье предложен выход на пакет прототипов системы фильтрации (СФ) служебных текстов (СТ).

Литературно-аналитический обзор по теме

Обзор сделан исходя из иерархии: система, подсистемы, блоки и т.д. Внутренняя структура систем фильтрации и алгоритмы их работы существенно зависят от природы фильтруемого материала. Так как цель данного исследования – развитие системы фильтрации текстовых сообщений, обзор посвящен рассмотрению аналогов в рамках систем семантической фильтрации [3÷4] с учетом настройки на специфику деятельности специалиста.

Согласно [5] система экстракции смысла из текстов, как и СФСТ, может состоять из достаточно большого количества подсистем: отбора ценной информации, автоматического понимания текста, построения тезаурусной онтологии понятий о работе специалиста, формирования когнитивных карт и т.д. Задача таких систем – извлечь необходимую (для сопоставления) информацию из поступающего текста методом наиболее близким к анализу текста самим специалистом, и при необходимости произвести ее окончательное отделение в соответствии с заданием.

По [3, 4, 6-10] основной метод сопоставления смыслов использует тезаурусы и рубрики заданного библиотечного классификатора (хотя бы примерно характеризующего целенаправленность смысла). Некоторые авторы [4, 5] уделяют внимание оценке пертинентности информации, при этом сводят определение пертинентности к сопоставлению рубрики текста с усреднённой рубрикой интересов специалиста или к снятию семантической неопределенности. В [3, 4, 6] отмечено о том, что специалиста не всегда интересуют все материалы по теме, а только их часть, поэтому необходимо разобраться с целеполаганием. При этом в качестве инструмента целеполагания предлагают рубрикационное сопоставление.

Состав СФ при условии выбора метода окончательной фильтрации (ранжирования), как базового, [4÷6] представлен в таблице 1.

Таблица 1

Пример структуры системы фильтрации информации

№ п/п	Подсистемы	Функции
1	настройки	учет специфики фильтруемого потока
2	отбора ценной информации	отбор релевантно-пертинентной информации по «образцу»
3	автоматического понимания текстов	извлечение и ранжирование терминов из поступающих текстов, определение семантических связей
4	автоматического рубрицирования текстов	определение рубрик, к которым может принадлежать текст и их возможное вероятностное оценивание
5	окончательной фильтрации	выделение релевантной и отсечение нерелевантной информации

Для подсистемы 1 общий подход изложен в [11]. А настройка на специфику деятельности специалистов медицинского учреждения и фильтруемого потока информации требует учета конкретики в соответствии с должностными обязанностями и компетенциями профильного специалиста.

Для подсистемы 2 основной метод извлечения информации об интересах пользователя – компьютерный диалог с ним на естественном языке [12]. В компьютере должна быть накоплена информация о предметной области, на основании которой он формирует вопрос (чаще в тестовой форме [12, 13]) и выводит его на экран для получения ответа специалиста. «Образец» для фильтрации – набор текстовых параметров (терминов, рубрик, семантических связей), а в качестве аналогов определения образца – это определение интересов пользователя. Задача отбора информации для тестирования заключается в формировании «Предварительного образца» по интересам пользователя для дальнейшей расстановки приоритетов. Состав подсистемы отбора информации для тестирования определяют в конце методом окончательной фильтрации [4,6], подразумевающей сопоставление информации по терминам, рубрикам и семантическим связям. Таким образом, в состав подсистемы должен входить отбор терминов и рубрик (в предположении возможности получения информации о семантических связях по терминам). В случае отбора рубрик достаточно, чтобы пользователь указал интересующую его библиотеку. Задача отбора терминов – их накопление из предметной области интересов пользователя. Таким образом, для отбора терминов выделяют 2 группы: основанные на словаре предметной области и использовании содержащихся в нем терминов [14], и основанные на интервьюировании субъекта [15-17]. В них генерируются вопросы пользователю, он развернуто отвечает на них и из ответов извлекаются термины.

Для подсистемы 3 в [18] приведён сравнительный обзор способов ранжирования терминов, оценивания их параметров и качества результатов, а также примеров программных реализаций; в [15] представлено комбинирование синтаксического, статистического и тематического ранжирования терминов. Цель в обоих случаях – вычислить вероятность вхождения термина в рубрики $P(T_k/G_l)$, где $P(\cdot)$ – вероятность заданного события, G_l – нахождение l-ой рубрики, T_k – использование -го термина при ее нахождении) при отсутствии автоматического рубрицирования. Попутно возможно дополнительное ранжирование.

Для подсистемы 4 найдено два аналога автоматического рубрицирования текстов: распознавание рубрик по [18] и СУБД ИРБИС64 [19]. Аналог [18] осуществляет количественную оценку характеристик текста (при помощи его автоматического понимания [6]) и последующего отнесения текста к определенной рубрике, рассматриваемой в данном случае как класс [20]. Важно отметить, что при этом используют сильно огрубленный список возможных рубрик, фактически назначаемый пользователем системы (в противном случае требуются слишком большие ресурсы для отнесения текста к рубрике). Аналог [19] позволяет получить для конкретного текста взвешенный список рубрик заданного библиотечного классификатора (УДК или ГРНТИ). Для этого должен быть задан словник рубрицируемого текста и указана библиотека, к которой требуется подключить систему (на сегодня реальна только ГПНТБ). Само рубрицирование основано на поиске текстов из указанной библиотеки, содержащих конкретный текущий термин и обнаружении рубрикационных отметок в найденных текстах.

Для подсистемы 5 найдены аналоги, основанные на использовании конкретных критериев отсечения информации, например, с помощью деревьев решений в задачах классификации Data Mining, оценки достоверности, целостности или непротиворечивости и т.п.

Материалы литературно-аналитического обзора позволили перейти к работе с аналогами.

Работа с аналогами

Исходя из значимости подсистем (см. табл.1) аналоги рассмотрены, начиная с подсистемы 5.

В таблицах 2÷5 приведено сравнение аналогов структурных составляющих СФСТ по выбранным критериям, помеченным (*), с весовой разбивкой.

Таблица 2

Сравнение аналогов окончательной фильтрации (подсистема 5)

№ п/п	Метод, взятый за основу аналога	Значения критерия сравнения: *)			∑
		полнота сопоставления тезаурусов (1/3)	полнота сопоставления целенатравленности (1/3)	возможность выхода на пертинентность (1/3)	
1	повышения пертинентности результатов информационного поиска в глобальной сети [4]	1	0,5	0,5	2/3
2	построения информационных семантических систем [3]	0,5	1	0	1/2
3	автоматического понимания текстов [6]	1	0,5	0,5	2/3
4	сопоставления результатов сравнения в сложном синтаксическом целом [7]	1	0	0	1/3
5	сопоставления отношений в простом предложении [8]	1	0	0	1/3
6	построения вероятностной модели смыслов [9]	1	0	0	1/3

*)

- полнота сопоставления тезауруса (0.0 – отсутствует; 0.5 – грубое; 1.0 – полное);
- полнота сопоставления целенаправленности (0.0 – отсутствует; 0.5 – грубое; 1.0 – полное);
- возможность выхода на пертинентность (0.0 – отсутствует; 0.5 – присутствует; 1.0 – учитывается).

По результатам сравнения отобраны аналоги метода повышения пертинентности результатов информационного поиска в глобальной сети [4] и автоматического понимания текстов [6]. Разумно в прототипе предложить их компиляцию.

Таблица 3

Сравнение аналогов автоматического рубрицирования текстов (подсистема 4)

№ п/п	Название аналога	Значения критерия сравнения: *)		∑
		полнота списка рубрик (1/2)	возможность неоднозначного взвешенного рубрицирования (1/2)	
1	Система распознавания рубрик [18]	0	0	0
2	Специализированная СУБД ИРБИС64 [19]	1/2	½	1

*)

- полнота списка рубрик (1 – полный; 0 – ограниченный);
- возможность неоднозначного взвешенного рубрицирования (1 – есть; 0 – нет).

По результатам сравнения наилучший аналог автоматического рубрицирования текстов – специализированная СУБД ИРБИС64.

Таблица 4

Сравнение аналогов дополнительного ранжирования терминов (подсистема 3)

№ п/п	Метод, взятый за основу	Значения критерия сравнения:*)			∑
		приоритетность функции дополнительного ранжирования терминов(1/3)	согласованность с субъектом(1/3)	простота внедрения или рациональность расчета $P(T_k)$ (1/3)	
1	тематического моделирования текстов на естественном языке [21]	0,5	0,5	0,25	5/12
2	комбинирования синтаксического, статистического и тематического ранжирования терминов [22]	0,5	0,5	0,25	5/12

*)

- приоритетность дополнительного ранжирования терминов (0 – не является приоритетным; 1 – является);
- согласованность с субъектом (0 – нет; 1 – высокая);
- простота внедрения в СФ информации из текстов или рациональность расчета $P(T_k)$ (0 – нерациональный расчет $P(T_k)$ и сложное внедрение; 1 – рациональный расчет $P(T_k)$ и простое внедрение).

Так как аналоги имеют одинаковые оценки, выбран компилятивный прототип, основанный на применении формулы:

$$P(T_k) = \sum_l P(G_l) * P(T_k/G_l), \quad (1)$$

где $P(T_k)$ – вероятность обнаружения k-ого термина, $P(G_l)$ – вероятность обнаружения l-ой рубрики, $P(T_k/G_l)$ – вероятность обнаружения k-ого термина при условии выбора l-ой рубрики.

Таблица 5

Сравнение аналогов отбора терминов (система 2)

№ п/п	Основа	Значения критерия сравнения:		Σ
		доступность для реализации (1/2)	интеллектуальность. (1/2)	
1	нахождение словаря предметной области [14]	0,125	0	0,125
2	провоцирование субъекта на выдачу текстов [15-17]	0,500	0,2	0,700

На основании результатов сравнения выбраны решения, основанные на провоцировании субъекта на выдачу текстов с последующим извлечением терминов.

Выход на пакеты прототипов

В результате исследования нами собран пакет научных прототипов 0-го и 1-го рангов (таблица 6) с критикой.

Таблица 6

Пакет научных прототипов системы фильтрации служебных текстов по семантической схожести с интересами субъекта

Ранг	Название прототипа	Ссылка	Критика
0	Система фильтрации	[4÷6, 15÷17, 19, 21, 22, 25÷27]	Системно-структурная неполнота
1	Подсистемы: -настройки на специфику фильтруемого материала, -отбора информации из текстов, -дополнительного ранжирования, -автоматического рубрицирования текстов, -окончательной фильтрации	 [11] [5, 15÷17] [21, 22] [19] [4, 6]	Функциональная неполнота

В связи с критикой этот пакет прототипов нуждается в доработке.

Прототипы младших рангов приведены в табл.7.

Таблица 7

Научные прототипы младших рангов

Ранг	Название прототипа	Ссылка	Критика
2	Блок дополнительного ранжирования терминов	[21, 22]	- нерациональность расчетов.
3	Узел отбора терминов	[15÷17]	- отсутствует функция кодирования терминов. - нерациональное извлечение терминов; - нерациональное накопление словника; - необъективный анализ результатов.
4	Центры: - накопления словника	[23]	- отсутствие блока проверки близости совокупности накопленных терминов к генеральной совокупности предметной области.
	- извлечения терминов из текстов	[15÷17]	- недоизвлечение терминов, которыми оперирует пользователь.
	- кодирования	[24]	- отсутствует настройка на кодирование словооснов терминов.
5	Модуль распознавания близости совокупности терминов к генеральной	[28]	- отсутствует настройка на информационную природу материала.

Также была проведена оценка корпоративного прототипа системы фильтрации, представленная в таблице 8 для МКМЦ «БОНУМ», г. Екатеринбург.

Таблица 8

Оценка корпоративного прототипа системы фильтрации текстов

Ранг	Название	Оценка качества фильтрации		Оценка временных затрат на фильтрацию 1 текста(мин)	
		фактическая	необходимая	фактическая	необходимая
0	Система фильтрации текстов	0,7	0,8	30	5
1	Подсистемы: - настройки на специфику, - отбора информации, - автоматического понимания, - дополнительного ранжирования, - окончательной фильтрации				

Таким образом, существующий корпоративный прототип в целом и по составляющим не удовлетворяет потребностей учреждения и нуждается в доработке. Эта доработка начата нами [29÷32]. Пояснения по пробелам в табл. 8: субъективная оценка качества подсистем невысока, однако требует отдельного объективного количественного рассмотрения.

Результаты и выводы:

- поставлена задача выхода на прототипы системы фильтрации текстов, релевантно-пертинентных деятельности специалиста медицинского учреждения,
- приведен материал литературно-аналитического обзора по теме, выявивший наличие 5-ти подсистем в структуре системы фильтрации текстов,
- составлен список авторитетных источников информации по теме поиска и обзора,
- проведен отбор аналогов системы фильтрации текстов и ее подсистем,
- предложен пакет научных и корпоративных прототипов системы фильтрации текстов,
- сделан вывод о возможном переходе к развитию прототипов на уровне моделирования.

Список литературы:

1. Грицюк Е.М. Анализ ситуации прерывания плановой работы при поступлении дополнительной информации, роль компьютерной поддержки / Е.М. Грицюк, С.Л. Гольдштейн, А.Г. Кудрявцев [и др.] // Системная интеграция в здравоохранении, №1, 2018, С. 5-17.
2. Гольдштейн С.Л. Фильтры текстов, отличающихся по смыслу от интересов пользователя. / Е.М. Грицюк, С.Л. Гольдштейн, А.Г. Кудрявцев [и др.] // НТИ, серия 2, Информационные процессы и системы, №2, 2019, С. 38-40.
3. Соломатин Н.М. Информационные семантические системы. –М: Высшая школа, 1989, - 127 с.
4. Савотченко С.Е. Современные аспекты повышения пертинентности результатов информационного поиска в глобальной сети / С.Е. Савотченко, Е.А. Проскурина.http://www.fundamentalresearch.ru/ru/article/view?id=34639
5. Goldshtein S.L. Modelling the extractor structure of the knowledge needed for fixing problematic situations involving a complicated object / S.L.Goldshtein, A.N.Averianova, E.M.Grisuk [and other]// Int. J. of Applied Systemic Studies, vol. 8, №2, P. 89-101).
6. Леонтьева Н.Н. Автоматическое понимание текстов: системы, модели, ресурсы, учеб. пособие, / Н.Н. Леонтьева. – М.: Издательский центр «Академия», 2006. – 306 с.
7. Пронька А.Г. Сопоставление результатов сравнения в сложном синтаксическом целом со сравнительно-количественными отношениями / А.Г. Пронька. – Волгоград: Известия ВГПУ, 2013. – 84 с.
8. Дмитриева Т.П. Сопоставительные и соотносительные отношения в простом предложении / Т.П. Дмитриева // Руснаука.http://www.rusnauka.com/16_ADEN_2011/Philologia /4_88667.doc.htm
9. Налимов В.В. Спонтанность сознания. Вероятностная теория смыслов и смысловая архитектоника личности / В.В. Налимов. – М.: Прометей, 1989. — 287 с.
10. Лихтциндер Б.Я. Вероятностные оценки уровня знаний учащихся / Б. Я. Лихтциндер, В.В. Пугин, Н.М.Татаринова // Конгресс конференций «Информационные технологии в образовании». http://ito.edu.ru/2008/Moscow/IX/IX-0-8028.html
11. Настройка корпоративных информационных систем на задачи предприятия / С.Л. Гольдштейн, И.В. Кашперский. – Екатеринбург: изд-во «Форт Диалог - Исеть», 2006. – 148 с.
12. Бабаева Ю.Д. Диалог с ЭВМ: психологические аспекты / Ю. Д. Бабаева, А. Е. Войскунский, В. В. Кобелев, О. К. Тихомиров// Вопросы психологии.1983,с. 25-35.
13. Ананченко И.В. Классификация компьютерных систем тестирования знаний учащихся // Международный журнал экспериментального образования. – 2016, № 4-2, С. 210-213. http://www.expeducation.ru/ru/article/view?id=9769.

14. Никифоров О.Ю. Использование адаптивных систем компьютерного тестирования // Гуманитарные научные исследования, 2014, № 4. http://human.snauka.ru/2014/04/6274
15. Леонтьев Д.А. Психология смысла: природа, строение и динамика смысловой реальности. http://lib.uni-dubna.ru/search/files/psy_leo_psy_smisla/1.pdf
16. Сутягин И. В. Методы формализации экспертных знаний для наполнения базы знаний // Молодой ученый, 2012,№1, Т.1, С. 151-153. https://moluch.ru/archive/36/4074/
17. Проскурин А.А. Автоматизированное извлечение теоретических знаний субъекта в рамках вопросно-ответного взаимодействия на основе анализа текста ответа на естественном языке. http://it-claim.ru/Persons/Proskurnin/StatyaProskurnin.pdf
18. Абрамов В.Е. Автоматическое рубрицирование и реферирование текстовой информации (в том числе на иностранных языках): дисс. … канд. техн. н. / В.Е. Абрамов. – М:, 2008.
19. Система ИР-БИС64.http://www.elnit.org/index.php?option=com_content&view=article&id=35&Itemid=108
20. Горелик А.Л. Методы распознавания: учеб. пособие / А.Л. Горелик, В.А. Скрипкин. – М: Высшая школа, 1984. – 208 с.
21. Коршунов А. Тематическое моделирование текстов на естественном языке / А.Коршунов, А.Гомзин // Труды Института системного программирования РАН, 2012.
22. Полушин В.В. Тематические модели для ранжирования рекомендаций текстового контента. ВКР. – М: МГУ им. М.В. Ломоносова, 2017.
23. Всеволодский С.Н. Архитектура интеллектуальной системы тестирования знаний с анализом ответов на естественном языке. / С.Н. Всеволодский, А.В. Гаврилов // Международная конференция ИСТ-2003 "Информационные системы и технологии", - Новосибирск, 2003, Т.3, С. 114-115.
24. Кодирование текста // Викиучебник. [2018—2018]. Дата обновления: 14.04.2018.https://ru.wikibooks.org/w/index.php?title=%D0%9A%D0%BE%D0%B4%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D1%82%D0%B5%D0%BA%D1%81%D1%82%D0%B0&oldid=145784
25. Адаптивные фильтры. /Под ред. К.Ф.Н.Коуэна и П.М.Гранта. – М: Мир, 1988. – 392 с.
26. Аксенова К. В. Фильтрация цифровых сигналов в режиме непрерывного поступления данных / К. В. Аксенова,В. П. Алексеев // Компьютерные исследования и моделирование. 2012, Т. 4, № 1, С. 55−61.
27. Пат. 2163032 Система адаптивной фильтрации аудиосигналов для улучшения разборчивости речи при наличии шума / Торбьерн В. СЕЛЬВЕ (US)
28. Теория вероятностей и математическая статистика : учебник для среднего профессионального образования / В. Е. Гмурман. - 12-е изд. - М: Изд. Юрайт, 2019. - 479 с.
29. Бызова А.К. Оценка точности классификации текстов в зависимости от их числа средствами Data Mining / А.К.Бызова, С.Л. Гольдштейн // Сб. «Физика. Технологии. Инновации. ФТИ-2015», - Екатеринбург: УрФУ, 2015, С. 224-229.
30. Бызова А.К. Развитие системы электронизации автоматизированного генератора системно-обоснованного технического задания путем разработки подсистемы структуризации/ А.К. Бызова, С.Л. Гольдштейн, Е.М. Грицюк // Системная интеграция в здравоохранении, 2016, №3, С. 5-24.
31. Бызова А.К. Подсистема структуризации вербальных текстов в АГ СО ТЗ, / А.К. Бызова, С.Л. Гольдштейн, Е.М. Грицюк // Свидетельство о госрегистрации программы для ЭВМ №2017616223 от 02.06.2017.
32. О снижении затрат времени на анализ и обработку текста специалистом с использованием компьютерной поддержки / Е.М.Грицюк, С.Л. Гольдштейн, Е.А. Дугина [и др.] // Системная интеграция в здравоохранении, №3, С. 22-32.