УДК 004.912

А.Д. Хомоненко, С.В. Логашев, С.А. Краснов

АВТОМАТИЧЕСКАЯ РУБРИКАЦИЯ ДОКУМЕНТОВ С ПОМОЩЬЮ ЛАТЕНТНО-СЕМАНТИЧЕСКОГО АНАЛИЗА И АЛГОРИТМА НЕЧЁТКОГО ВЫВОДА МАМДАНИ

Хомоненко А.Д., Логашев С.В., Краснов С.А. **Автоматическая рубрикация документов с помощью латентно-семантического анализа и алгоритма нечёткого вывода Мамдани.**

Аннотация. Предлагается подход к автоматической рубрикации текстовых документов на основе совместного применения метода латентно-семантического анализа (ЛСА) и алгоритма нечёткого вывода Мамдани. Метод ЛСА используется для смыслового анализа информации в системах электронного документооборота путем выявления семантических зависимостей между термами документов и получения коэффициента соответствия сравниваемых векторов.

Предлагается база правил для алгоритма нечёткого вывода Мамдани, реализующего автоматическую рубрикацию документов по множеству заданных тематик с возможностью автоматизированного контроля за распределением документов не соответствующим заданным тематикам или имеющим сходство сразу по нескольким тематическим категориям на основе результатов латентно-семантического анализа.

Ключевые слова: рубрикация документов, нечеткий вывод, латентно-семантический анализ, база правил, алгоритм нечёткого вывода Мамдани.

Khomonenko A.D. Logashev S.V., Krasnov S.A. **Automatic Categorization of Documents Using Latent Semantic Analysis and Fuzzy Inference Algorithm of Mamdani.**

Abstract. We propose an approach to the automatic categorization of text documents based on the joint application of the method of latent semantic analysis (LSA) and fuzzy inference Mamdani algorithm. Method LSA is used for the semantic analysis of information in electronic document management systems by identifying semantic relationships between terms of documents and receipt of the compliance rate of the compared vectors. The rule base is proposed for fuzzy inference algorithm of Mamdani implementing the automatic rubrication of documents for a variety of given topics enabling automated monitoring of the distribution of documents not relevant to the specified topics, or having similarities in several thematic categories on the basis of the results of latent semantic analysis.

Keywords: rubrication of documents; fuzzy inference; latent semantic analysis; the rule base; a fuzzy inference Mamdani algorithm.

1. Введение.

Целью работы является выработка подхода к решению задачи автоматической рубрикации документов по заданным тематическим рубрикам [1, 2]. Для этого предлагается использовать совместно метод латентно-семантического анализа и алгоритм нечёткого вывода Мамдани, что определяет новизну предлагаемого подхода.

Для решения задачи автоматической рубрикации документов используются методы семантического анализа и автоматического разделения поступающей информации по заданным рубрикам.

В последние годы в задачах автоматической рубрикации, все больше внимания привлекают современные подходы к семантическому анализу, обеспечивающие лучшее качество [3–6].

Выявление семантической структуры при помощи латентно-семантического анализа выполняется алгоритмически и не требует ручного составления словарей. Однако результатом применения метода ЛСА является численное значение вероятности совпадения сравниваемых документов, что не позволяет полностью автоматически рубрицировать документ. В частности, широко используемые классические методы (основанные на базах знаний или машинном обучении) позволяют рубрицировать текстовые документы с рядом ошибок.

Для повышения точности решения рассматриваемой задачи, предлагается использовать методы и алгоритмы нечеткого вывода, т.к. в решении задачи автоматической рубрикации документов имеется нечеткость значений анализируемых параметров.

В основе реализации методов нечеткого вывода лежит теория нечетких множеств и основанная на ней нечеткая логика. В статье предлагается модель рубрикации документов, основанная на алгоритме нечеткого вывода. Для реализации построенной модели использовалась среда MATLAB и специальный пакет расширения Fuzzy Logic Toolbox [7]. Широкая область использования и корректность результатов позволяют применить математический аппарат нечеткого вывода при моделировании сложных процессов в области автоматической рубрикации [8, 9].

Применение метода нечёткого вывода позволяет создать метод автоматической рубрикации документов с возможностью автоматизированного принятия решения по полисемантическим документам и повысить уровень автоматизации процесса рубрикации документов.

2. Применение метода ЛСА для установления степени близости документов. Метод латентно-семантического анализа позволяет автоматически проанализировать содержимое текстовой информации, содержащейся в документах, и выявлять скрытые семантические (смысловые) связи между документами [10].

Исходной информацией для ЛСА является матрица А терм-документ, которая описывает используемый для обучения системы набор документов. Элементы этой матрицы содержат частоты использования каждого термина в каждом документе [5, 6].

Следующим шагом является разложение полученной матрицы терм-документ. Согласно теореме о сингулярном разложении, любая вещественная прямоугольная матрица A может быть разложена в произведение трёх матриц [11, 12, 13]:

$$A = VWU^T. \quad (1)$$

На последнем этапе необходимо рассчитать степень соответствия векторов (документов), обычно для этого используется математическая операция скалярного произведения векторов.

В качестве примера рассмотрим обучающую выборку документов D, которая автоматически разделена методом ЛСА [5, 14] на две различные группы документов da1-da5 и db1-db3 (таблицы 1 и 2).

В группе da1-da5 речь идёт об указах председателя Правительства Российской Федерации. В группе db1-db3 речь идёт о федеральных законах. Документ db4 — исходный документ для проведения рубрицирования.

Таблица 1. Состав документов группы a

Группа a:	
1.	*Указ Председателя правительства Российской Федерации* от 26 августа 2010 г. N 1110 г. *Москва* "Об установлении *ежемесячной надбавки* за важность выполняемых задач специалистам *физической подготовки*"
2.	*Указ Председателя правительства Российской Федерации* от 30 сентября 2010 г. N 1280 г. *Москва* "О предоставлении *госслужащим* жилых помещений по договору социального найма и служебных помещений"
3.	*Указ Председателя правительства* от 26 августа 2010 г. N 1115 г. *Москва* "Об установлении *ежемесячной надбавки госслужащим*, проходящим военную службу по контракту, за квалификационный уровень *физической подготовленности.*
4.	*Указ Председателя правительства Российской Федерации* от 24 апреля 2010 г. N 100 г. *Москва* "Об утверждении Инструкции об условиях и порядке приема в учреждения высшего профессионального образования"
5.	*Указ Председателя правительства Российской Федерации* от 29 марта 2010 г. N 299 г. *Москва* «О Порядке проведения в *Правительстве Российской Федерации под руководством председателя* антикоррупционной экспертизы нормативных *правовых* актов».

Таблица 2. Состав документов группы b

Группа b:	
1.	*Дмитрий Медведев подписал Федеральный закон* «О *внесении изменений* в *Федеральный закон* «О *федеральном* бюджете на 2010 год и на плановый период».
2.	Президент *подписал Федеральный конституционный закон* «О *внесении изменений* в *Федеральный конституционный закон* «О *Конституционном* Суде Российской Федерации».
3.	*Дмитрий Медведев подписал Федеральный закон* «О *внесении изменений* в статьи 14 и 15 *Федерального закона* «О политических партиях».
4.	*Дмитрий Медведев подписал* распоряжение о проведении Международного общественного форума «Роль народной дипломатии в развитии международного гуманитарного *физического* сотрудничества, и *права военнослужащих*».

Результаты распределения групп документов представлены в таблице 3. Для оценки соответствия документов группам рассмотрим максимальные полученные значения элементов матриц полученных с помощью скалярного произведения векторов, далее по тексту косинусоидальная мера близости [5], которая позволяет получить коэффициент сходства между различными векторами (рисунок 1).

Таблица 3. Значения аппроксимирующей матрицы X

	a1	a2	a3	a4	a5	b1	b2	b3	b4
a1		6.665	9.276	6.623	11.315	0.748	1.078	1.363	1.152
a2			7.943	5.697	9.728	0.359	0.642	0.908	0.926
a3				7.864	13.448	1.536	1.929	2.219	1.515
a4					9.731	-0.393	-0.110	0.212	0.756
a5						-0.348	0.135	0.660	1.362
b1							16.662	15.406	3.746
b2								15.443	3,710
b3									3.551
b4									

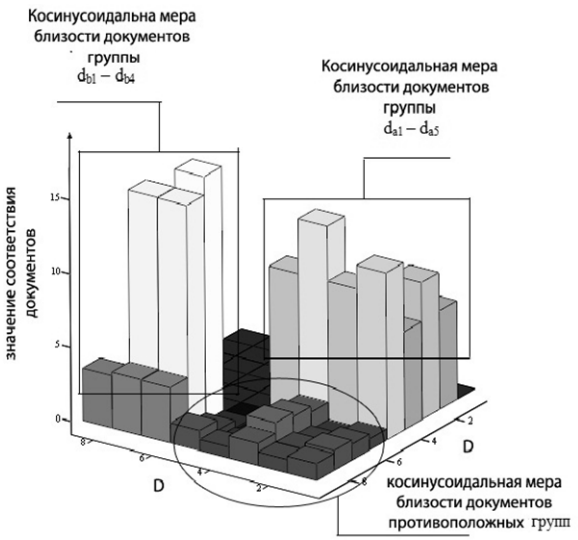

Рис. 1. Гистограмма значений аппроксимирующей матрицы X

Результатом применения метода ЛСА при сравнении документа b4 с остальными документами выборки является таблица значений. Она содержит значения входных переменных для алгоритма Мамдани (таблица 4).

Таблица 4. Исходные данные для нечёткого вывода

Название функции	Значение функции
b4-a1	1.152
b4-a2	0.926
b4-a3	1.515
b4-a4	0.756
b4-a5	1.362
b4-b1	3.746
b4-b2	3,710
b4-b3	3.551

Рассмотрим решение задачи автоматической рубрикации по результатам, полученным методом ЛСА, с помощью алгоритма нечеткого вывода Мамдани. С этой целью, прежде всего, сформируем соответствующую базу правил.

3. База правил для рубрикации документов с помощью алгоритма Мамдани. Рубрикация рассматриваемого текста проводится по значениям столбца аппроксимирующей матрицы X. В качестве текущего рубрицируемого документа возьмём документ b4. Введём 8 нечетких лингвистических переменных B4-A1, B4-A2, B4-A3, B4-A4, B4-A5, B4-B1, B4-B2, B4-B3. Каждая переменная получает соответствующее значение из матрицы X. Результат работы — значения лингвистических переменных, формируемых на основании уровня соответствия текущего документа к группам A и B.

Лингвистическая переменная определяется как кортеж:

<?, T, X, G, M>, где:

? — наименование или название лингвистической переменной;

T — базовое терм-множество лингвистической переменной или множество ее значений (термов);

X — область определения (универсум) нечетких переменных, которые входят в определение лингвистической переменной ?;

G — синтаксическая процедура, описывающая процесс образования новых термов;

M — семантическая процедура образования новых термов.

Определим терм-множества входных лингвистических переменных:

1) {B4-A1 = Уровень соответствия документа B_4 документу A_1, T = {низкий, средний, высокий}, X=[0;17]};

2) {B4-A2 = Уровень соответствия документа B_4 документу A_2, T = {низкий, средний, высокий}, X=[0;17]};

3) {B4-A3 = Уровень соответствия документа B_4 документу A_3, T = {низкий, средний, высокий}, X=[0;17]};

4) {B4-A4 = Уровень соответствия документа B_4 документу A_4, T = {низкий, средний, высокий}, X=[0;17]};

5) {B4-A5 = Уровень соответствия документа B_4 документу A_5, T = {низкий, средний, высокий}, X=[0;17]};

6) {B4-B1 = Уровень соответствия документа B_4 документу B_1, T = {низкий, средний, высокий}, X=[0;17]};

7) {B4-B2 = Уровень соответствия документа B_4 документу B_2, T = {низкий, средний, высокий}, X=[0;17]};

8) {B4-B3 = Уровень соответствия документа B_4 документу B_3, T = {низкий, средний, высокий}, X=[0;17]};

А так же определим терм-множества выходных лингвистических переменных:

9) {Группа_А = Уровень соответствия документа B_4 группе А, Т = {низкий_уровень_соответствия, эксперт, высокий_уровень_ соответствия}, X=[0;1]}.

10) {Группа_В = Уровень соответствия документа B_4 группе В, Т = {низкий_уровень_соответствия, эксперт, высокий_уровень_ соответствия}, X=[0;1]}.

В редакторе MATLAB Fuzzy Logic Toolbox задаются функции принадлежности всех входных и выходной лингвистической переменных. Значения функций принадлежности определяются, учитывая мнения экспертов.

После создания лингвистических переменных, опираясь на знания экспертов, сформируем базу правил системы нечеткого вывода следующего вида:

1. If (B4-A1 is средний) and (B4-A2 is средний) and (B4-A3 is низкий) and (B4-A4 is низкий) and (B4-A5 is низкий) and (B4-B1 is низкий) and (B4-B2 is низкий) and (B4-B3 is низкий) then (Группа_А is низкий_уровень_соответствия) (Группа_В is низкий_уровень_соответствия)

2. If (B4-A1 is высокий) and (B4-A2 is высокий) and (B4-A3 is высокий) and (B4-A4 is высокий) and (B4-A5 is высокий) and (B4-B1 is низкий) and (B4-B2 is низкий) and (B4-B3 is низкий) then (Группа_А is высокий_уровень_соответствия) (Группа_В is низкий_уровень_соответствия)

3. If (B4-A1 is низкий) and (B4-A2 is низкий) and (B4-A3 is низкий) and (B4-A4 is низкий) and (B4-A5 is низкий) and (B4-B1 is высокий) and (B4-B2 is высокий) and (B4-B3 is высокий) then (Группа_А is низкий_уровень_соответствия) (Группа_В is высокий_уровень_соответствия)

4. If (B4-A1 is средний) and (B4-A2 is средний) and (B4-A3 is средний) and (B4-A4 is средний) and (B4-A5 is средний) and (B4-B1 is средний) and (B4-B2 is средний) and (B4-B3 is средний) then (Группа_А is эксперт) (Группа_В is эксперт)

5. If (B4-A1 is высокий) and (B4-A2 is высокий) and (B4-A3 is высокий) and (B4-A4 is высокий) and (B4-A5 is высокий) and (B4-B1 is высокий) and (B4-B2 is высокий) and (B4-B3 is высокий) then

(Группа A is высокий_уровень_соответствия) (Группа_B is высокий_уровень_соответствия)

...

19. If (B4-B1 is высокий) and (B4-B2 is высокий) and (B4-B3 is средний) then (Группа_B is высокий_уровень_соответствия)

Каждая входная переменная получает значение, образованное после применения ЛСА для документа b4 и остальных рассматриваемых документов (рисунок 2).

Из рисунка 2 видно, что после применения алгоритма нечёткого вывода Мамдани выходные значения результирующих функций примут значения: Группа_A=0.157; Группа_B=0.2.

Из приведённых ниже графиков принадлежности лингвистических переменных Группа_A и Группа_B видно, что значения переменных обоих функций соответствуют лингвистическому терму «низкий_уровень_соответствия». Значение этого терма является результатом работы алгоритма нечёткого вывода. Документ не отнесён ни к одной из групп.

Если значения входных функций примут значения, указанные в таблице 5, то результирующие значения функций алгоритма нечёткого вывода изменятся и примут вид, показанный на рисунке 3.

Таблица 5. Исходные данные для алгоритма нечёткого вывода

Название функции	Значение функции
b4-a1	4.1
b4-a2	13.5
b4-a3	14.9
b4-a4	8.5
b4-a5	8.5
b4-b1	8.5
b4-b2	8.5
b4-b3	3.8

В приведенном примере Группа_A=0.8, Группа_B=0.498. Эти значения соответствуют лексическим термам Группа_A=«высокий_уровень_соответствия», Группа_B=«эксперт». В этом случае полисемантический документ автоматически относится к группе документов А, а возможность его записи в группу Б определяется оператором.

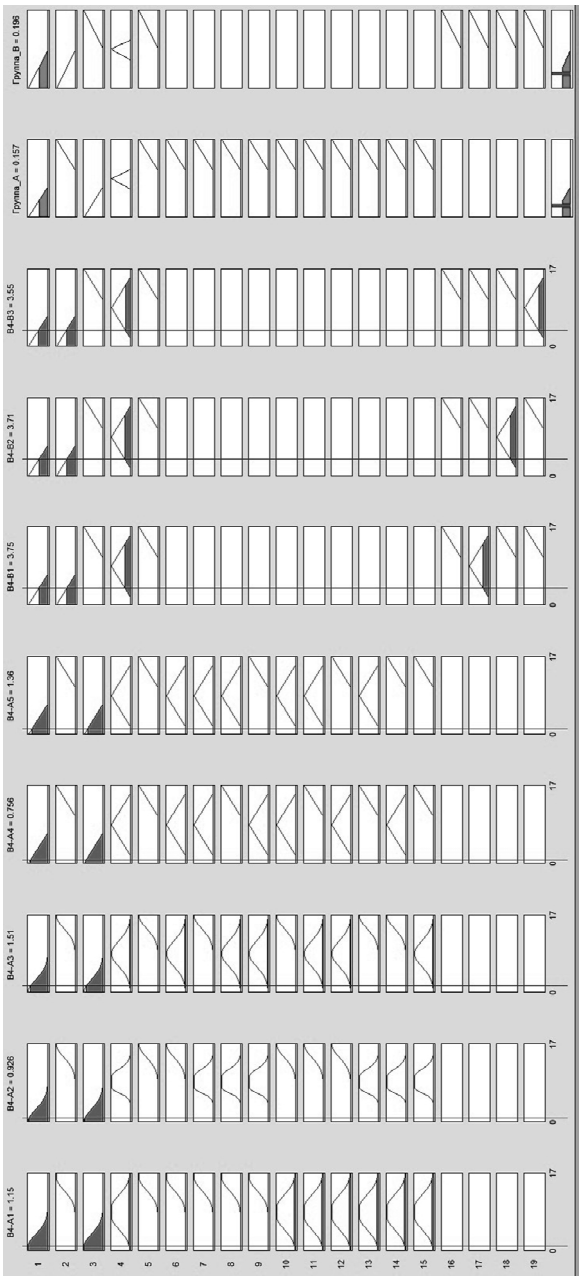

Рис. 2. Состав правил и вид решения при несоответствии документов

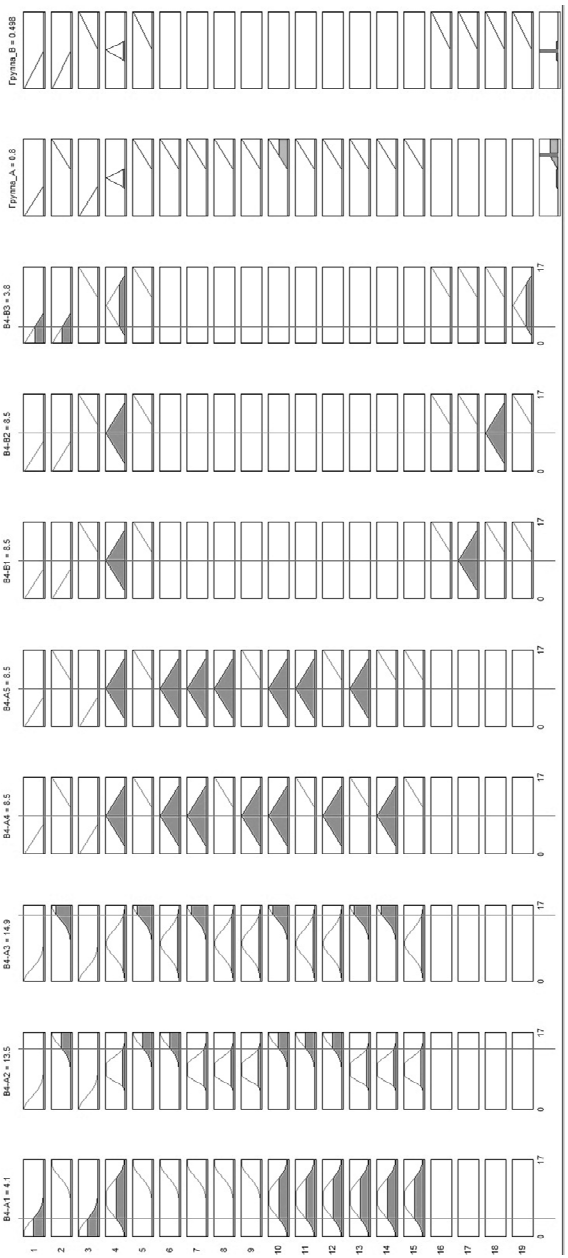

Рис. 3. Состав правил и вид решения при соответствии документов

4. Заключение. В настоящей статье предложен метод автоматической рубрикации текстовых документов на основе совместного применения метода латентно-семантического анализа (ЛСА) и алгоритма нечёткого вывода Мамдани.

Использование метода ЛСА и алгоритма нечёткого вывода с предложенной базой правил для алгоритма нечёткого вывода Мамдани позволяет автоматически принимать решение по рубрикации текстовых документов, что позволяет снизить временные затраты на рубрикацию. При этом не возрастает количество семантических ошибок рубрикации вследствие применения метода ЛСА с использованием набора ключевых слов. В ситуациях, когда возможно появление ошибок, система рубрикации предлагает оператору рубрицировать документ.

Дальнейшие исследования целесообразно продолжить в направлении оценивания оперативности и эффективности применения предложенного подхода, например, на основе вероятностных моделей, таких как [15]. Кроме того, в направлении обоснования выбора алгоритма нечеткого вывода [16-18] по соотношению трудоемкости и точности получаемых решений.

Литература

1. *Агеев М.С., Добров Б.В., Лукашевич Н.В.* Автоматическая рубрикация текстов: методы и проблемы // Учебные записки Казанского государственного университета. Физико-математические науки. 2008. Вып. 150. № 4. С. 25–40.
2. *Гареев А. Ф.* Автоматическое тематическое рубрицирование сообщений средств массовой информации на основе применения технологии нейронных сетей // Информационные технологии. 1999. № 5. С. 26–33.
3. *Papka R., Allan J.* Document classification using multiword features // Proceedings of the A CM International Conference on Information and Knowledge Management (CIKM-98). New York. ACM Press. 1998. pp. 124–131.
4. *Manning C.D., Raghavan P., Schütze H.* An Introduction to Information Retrieval Draft. Online edition // Cambridge University Press. 2009. 544 p.
5. *Хомоненко А.Д., Краснов С.А.* Применение метода латентно-семантического анализа для автоматической рубрикации документов // Известия Петербургского университета путей сообщения. 2012. № 2(31). С. 124–132.
6. *Бубнов В.П. и др.* Модели информационных систем: учеб. пособие // М.: ФГБОУ «Учебно-методический центр по образованию на железнодорожном транспорте». 2015. 188 с.
7. *Mamdani E.H.* Application of fuzzy logic to approximate reasoning using linguistic Systems // Fuzzy Sets and Systems. 1977. vol. 26. pp. 1182–1191.
8. *Войцеховский С.В., Хомоненко А.Д.* Выявление вредоносных программных воздействий на основе нечеткого вывода // Проблемы информационной безопасности. Компьютерные системы. 2011. № 3. С. 81–91.
9. *Хомоненко А.Д., Войцеховский С.В., Логашев С.В., Дашонок В.Л.* Устранение семантических противоречий в elibrary.ru на основе нечёткого вывода // Проблемы информационной безопасности. Компьютерные системы. 2015. № 1. С. 24–33.

10. *Хомоненко А.Д., Дашонок В.Л., Краснов С.А.* Выявление противоречий в семантически близкой информации на основе латентно-семантического анализа // Проблемы информационной безопасности. Компьютерные системы. 2014. № 2. С. 73–84.
11. *Foltz P.W.* Using latent semantic indexing for information filtering // In ACM Conference on Office Information Systems (COIS). 1990. pp. 40–47.
12. *Dumais S.* Latent semantic indexing: TREC-3 report // In Proc. of the Third Text REtrieval Conference. 1995. pp. 219–230.
13. *Landauer T., Foltz P. and Laham D.* An introduction to Latent Semantic Analysys // Discourse processes. 1998. vol. 25. no. 2–3. С. 259–284.
14. *Кураленок И.Е., Некрестьянов И.С.* Автоматическая классификация документов на основе латентно-семантического анализа // Труды первой всероссийской научно-методической конференции "Электронные библиотеки: перспективные методы и технологии, электронные коллекции". СПб. 1999. С. 89–96.
15. *Хомоненко А.Д., Краснов С.А., Еремин А.С.* Оценка оперативности автоматической рубрикации документов с помощью модели нестационарной системы обслуживания с эрланговским распределением длительности интервалов между запросами // Проблемы информационной безопасности. Компьютерные системы. 2012. № 3. С. 14–21.
16. *Takagi T., Sugeno M.* Fuzzy Identification of Systems and Its Applications to Modeling and Control // IEEE Trans. Systems, Man, and Cybernetics. 1985. vol. 15. no. 1. pp. 116–132.
17. *Леоненков А.* Нечеткое моделирование в среде MATLAB и fuzzyTECH // СП: БХВ-Петербург. 2003. 736 с.
18. *Штовба С.Д.* Проектирование нечётких систем средствами MATLAB // М.: Горячая линия-Телеком. 2007. 288 с.

References

1. Ageev M.S., Dobrov B.V., Lukashevich N.V. [Automatic categorization of texts: methods and problems.] *Uchebnye zapiski Kazanskogo gosudarstvennogo universiteta. Fiziko-matematicheskie nauki – Scientific notes of the Kazan State University. Physics and mathematics*. 2008. Issue 150. vol. 4. pp. 25–40. (In Russ.).
2. Gareev A. F. [Thematic automatically classifying messages of the media through the use of neural network technology.]. *Informacionnye tehnologii – Information Technology*. 1999. vol. 5. pp. 26–33. (In Russ.).
3. Papka R., Allan J. Document classification using multiword features. Proceedings of the A CM International Conference on Information and Knowledge Management (CIKM-98). New York. ACM Press. 1998. pp. 124–131.
4. Manning C.D., Raghavan P., Schütze H. An Introduction to Information Retrieval Draft. Online edition. Cambridge University Press. 2009. 544 p.
5. Khomonenko A.D., Krasnov S.A. [Application of the method of latent semantic analysis for automatic categorization of documents]. *Izvestija Peterburgskogo universiteta putej soobshhenija – Proceedings of St. Petersburg State University of Communication.* 2012. vol. 2(31). pp. 124–132. (In Russ.).
6. Bubnov V.P. *Modeli informacionnyh sistem: ucheb. posobie* [Models of information systems: proc. Manual]. Moscow: FGBOU «Uchebno-metodicheskij centr po obrazovaniju na zheleznodorozhnom transporte». 2015. 188 p. (In Russ.).
7. Mamdani E.H. Application of fuzzy logic to approximate reasoning using linguistic Systems. Fuzzy Sets and Systems. 1977. vol. 26. pp. 1182–1191.
8. Vojcehovskij S.V., Khomonenko A.D. [The revealing of harmful program influences on the basis of fuzzy inference] *Problemy informacionnoj bezopasnosti.*

Komp'juternye sistemy – Problems of information security. Computer systems. 2011. vol. 3. pp. 81–91. (In Russ.).
9. Khomonenko A.D., Vojcehovskij S.V., Logashev S.V., Dashonok V.L. [Resolving semantic inconsistencies in elibrary.ru on the basis of fuzzy inference]. *Problemy informacionnoj bezopasnosti. Komp'juternye sistemy – Problems of information security. Computer systems.* 2015. vol 1. pp. 24–33. (In Russ.).
10. Khomonenko A.D., Dashonok V.L., Krasnov S.A. [The identification of contradictions in semantically close information based on latent semantic analysis]. *Problemy informacionnoj bezopasnosti. Komp'juternye sistemy – Problems of information security. Computer systems.* 2014. vol. 2. pp. 73–84. (In Russ.).
11. Foltz P.W. Using latent semantic indexing for information filtering. In ACM Conference on Office Information Systems (COIS). 1990. pp. 40–47.
12. Dumais S. Latent semantic indexing: TREC-3 report. In Proc. of the Third Text REtrieval Conference. 1995. pp. 219–230.
13. Landauer T., Foltz P., Laham D. An introduction to Latent Semantic Analysys. Discourse processes. 1998. vol. 25. no. 2–3. pp. 259–284.
14. Kuralenok I.E., Nekrest'janov I.S. [Automatic categorization of documents based on latent semantic analysis] *Trudy pervoj vserossijskoj nauchno-metodicheskoj konferencii "Jelektronnye biblioteki: perspektivnye metody i tehnologii, jelektronnye kollekcii"*. SPb [Proceedings of the First All-Russian Scientific Conference "Digital Libraries: Advanced Methods and Technologies, Digital Collections"]. 1999. pp. 89–96. (In Russ.).
15. Khomonenko A.D., Krasnov S.A., Eremin A.S. [Evaluation of the efficiency of the automatic rubrication of documents using the model of non-stationary queueing systems with erlangovsky distribution of duration of intervals between requests] *Problemy informacionnoj bezopasnosti. Komp'juternye sistemy – Problems of information security. Computer systems.* 2012. vol. 3. pp. 14–21. (In Russ.).
16. Takagi T., Sugeno M. Fuzzy Identification of Systems and Its Applications to Modeling and Control. *IEEE Trans. Systems, Man, and Cybernetics*. 1985. vol. 15. no. 1. pp. 116–132.
17. Leonenkov A. *Nechetkoe modelirovanie v srede MATLAB i fuzzyTECH* [Fuzzy modeling in MATLAB and fuzzyTECH]. SPb.: BHV-Peterburg, 2003. 736 p. (In Russ.).
18. Shtovba S.D. *Proektirovanie nechjotkih sistem sredstvami MATLAB* [The design of fuzzy systems by means of MATLAB]. Moscow: Gorjachaja linija-Telekom. 2007. 288 p. (In Russ.).

Хомоненко Анатолий Дмитриевич — д-р техн. наук, профессор, заведующий кафедрой информационных и вычислительных систем, ФГБОУ ВПО Петербургский государственный университет путей сообщения Императора Александра I. Область научных интересов: численная теория массового обслуживания, программирование, операционные и информационные системы. Число научных публикаций — 150. khomon@mail.ru, http://www.pgups.ru; Московский пр., 9, Санкт-Петербург, 190031; р.т.: 457-80-23, Факс: 310-75-25.

Khomonenko Anatoly Dmitrievich — Ph.D., Dr. Sci., professor, head of information and computing systems department, Petersburg State Transport University. Research interests: queuing systems, artificial intelligence, databases. The number of publications — 150. khomon@mail.ru, http://www.pgups.ru; 9, Moskovsky pr., Saint Petersburg, 190031; office phone: 457-80-23, Fax: 310-75-25.

Логашев Сергей Вячеславович — преподаватель кафедры, Военно-космическая академия имени А.Ф. Можайского. Область научных интересов: базы данных, информационные системы, системы поддержки принятия решения. Число научных публикаций — 2. loga1977@yandex.ru; ул. Ждановская 13, Санкт-Петербург, 197198; р.т.: 8-906-225-67-76.

Logashev Sergej Vjacheslavovich — teacher, Mozhaisky Military Space Academy. Research interests: databases, information systems, systems of support of decision-making. The number of publications — 2. loga1977@yandex.ru; 13, Zhdanovskaya street, St.-Petersburg, 197198, Russia; office phone: 8-906-225-67-76.

Краснов Сергей Александрович — к-т техн. наук, старший преподаватель, Военно-космическая академия имени А.Ф. Можайского. Область научных интересов: информационные технологии, защита информации, системы искусственного интеллекта. Число научных публикаций — 30. kras25@rambler.ru; ул. Ждановская 13, Санкт-Петербург, 197198; р.т.: 89117346550.

Krasnov Sergey Aleksandrovich — senior lecturer, Mozhaisky Military Space Academy. Research interests: information technology, information security, artificial intelligence systems. The number of publications — 30. kras25@rambler.ru; 13, Zhdanovskaya street, St.-Petersburg, 197198, Russia; office phone: 89117346550.

РЕФЕРАТ

Хомоненко А.Д., Логашев С.В., Краснов С.А. **Автоматическая рубрикация документов с помощью латентно-семантического анализа и алгоритма нечёткого вывода Мамдани.**

Целью работы является выработка подхода к решению задачи автоматической рубрикации документов по заданным тематическим рубрикам. Для этого предлагается использовать совместно метод латентно-семантического анализа и алгоритм нечёткого вывода Мамдани, что определяет новизну предлагаемого подхода. Выявление семантической структуры при помощи латентно-семантического анализа выполняется алгоритмически и не требует ручного составления словарей. Для повышения точности решения рассматриваемой задачи, предлагается использовать методы и алгоритмы нечеткого вывода, т.к. в решении задачи автоматической рубрикации документов имеется нечеткость значений анализируемых параметров. Для реализации построенной модели использовалась среда MATLAB и специальный пакет расширения Fuzzy Logic Toolbox. Исходной информацией для ЛСА является матрица А терм-документ, которая описывает используемый для обучения системы набор документов. Далее производим разложение полученной матрицы терм-документ и рассчитываем степень соответствия векторов (документов), обычно для этого используется математическая операция скалярного произведения векторов. Полученные значения обрабатываются алгоритмом нечёткого вывода Мамдани, результатом которого являются значения выходных функций о принадлежности документа каждой из существующих рубрик.

SUMMARY

Khomonenko A.D., Logachev S.V., Krasnov S.A. **Automatic Categorization of Documents Using Latent Semantic Analysis and Fuzzy Inference Algorithm of Mamdani.**

The aim of this work is to develop an approach to the problem of automatic categorization of documents by given subject headings. For this purpose, it is proposed to use jointly the method of latent semantic analysis and fuzzy inference algorithm of Mamdani, which determines the novelty of the proposed approach. Identifying semantic patterns using latent semantic analysis is performed algorithmically and does not require manual compilation of dictionaries. To improve the accuracy of the considered problem, it is proposed to use methods and algorithms of fuzzy inference, as in the solution of the problem of the automatic rubrication of documents there are ambiguities in the values of the analyzed parameters. For the implementation of the constructed model, MATLAB and a special expansion pack Fuzzy Logic Toolbox were used. Source data for LSA is a matrix A term-document that describes a set of documents being used to train the system. Next, we make the decomposition of the resulting matrix of term-document and calculate the degree of correspondence of vectors (documents), usually via a mathematical operation of scalar product of vectors. The obtained values are processed by fuzzy inference algorithm of Mamdani, resulting values are output functions of a document belonging to each of the existing headings.