Фоминых С. В. Автоматический независимый от языка анализ авторства патристических текстов на основании статистики частот переходов // Исторический журнал: научные исследования. – № 5. – 2017. – С. 70-79. DOI: 10.7256/2454-0609.2017.5.23699.

e-mail: thominykh.sv@gmail.com

Автоматический независимый от языка анализ авторства патристических текстов на основании статистики частот переходов

Abstract. Описывается применение алгоритма независимого от языка автоматического анализа авторства на основании подсчета статистики частот переходов двухбуквенных сочетаний для патристических авторов со II по XII век писавших на древнегреческом языке. Авторство анализируемого текста определяется на основании близости по заданному расстоянию к эталонным текстам различных авторов. Расстояние подсчитывается на основе частот переходов одних двухбуквенных сочетаний в другие. Обсуждается зависимость точности алгоритма от величины анализируемого текста и эталонных текстов каждого из предполагаемых авторов. Тестирование проводится методом повторяющегося скользящего контроля по k-блокам и отдельно на 35 выбранных текстах 3-х авторов. Предлагается модификация алгоритма в некоторых случаях дающая лучший результат на тестовых данных. Делается вывод о достаточно высокой точности описанного алгоритма и о возможности его применения для решения реальных задач. В качестве примера использования описывается определение авторства текстов De creatione hominis sermo 1, 2 (CPG 3215, 3216) между свт. Василием Великим и свт. Григорием Нисским.

Ключевые слова. Василий Великий, Григорий Нисский, определение авторства, передача текста, обработка естественного языка, вычислительная статистика, статистический анализ, Византийские исследования, патристика, историческая информатика.

Keywords. Basil the Great, Gregory of Nyssa, authorship attribution, textual transmission, natural language processing, computational statistics, statistical analysis, Byzantine studies, patristics, digital history.

Постановка задачи

Задача автоматического определения авторства может рассматриваться как частный случай задачи классификации по прецедентам. Данный класс задач решается сравнением анализируемого объекта с эталонами (прецедентам) по некоторым признакам. Нахождение возможности выделения из анализируемых текстов признаков без интерпретации языка, на котором эти тексты написаны, означает, что исследователь, вообще говоря, может не знать используемый в текстах язык.

Пусть есть $2 \ldots N$ авторов, имеющих соответствующие им эталонные тексты. Под задачей идентификации авторства будем понимать автоматическое определение наиболее вероятного автора из заданных N вариантов для некоторого текста, не входящего ни в один из эталонных текстов. Под эталонным текстом не обязательно понимается одно законченное произведение, он может состоять из соединения нескольких текстов данного автора или лишь части одного сочинения.

Описание методики

Предлагаемый метод основан на анализе близости распределений частот переходов одних двухбуквенных сочетаний в другие (признаки) в анализируемом тексте и в эталонных текстах каждого автора. Данный подход был изложен в 2000 г. Д. В. Хмелевым [7]. Ранее подсчет частоты двухбуквенных сочетаний (но не их переходов) для определения авторства предлагался В. Беннеттом в 1976 г. [14, p. 257]. Д. В. Хмелев, однако, основывался на исследовании А. А. Маркова начала XX в. [7]. В последствии подходы, основанные на обработке буквенных последовательностей (в широком смысле), были признаны одним из наиболее эффективных способов для определения авторства [16; 18].

Пусть имеется текст, состоящий из пробелов и символов некоего алфавита. Подсчитаем в этом тексте количество последовательных переходов одних двухбуквенных сочетаний в другие. В качестве примера приведем начальные слова 1-го слова свт. Григория Богослова: “ἀναστάσεως ἡμέρα, καὶ ἡ ἀρχὴ δεξιά”. Количество переходов двухбуквенных сочетаний в этом тексте предлагается считать следующим образом: ἀν → ασ = 1; ασ → τά = 1; τά → σε = 1; σε → ως = 1; ἡμ → έρ = 1; ἀρ → χὴ = 1; δε → ξι = 1.

Имея данные о таких переходах для всего анализируемого текста, можно провести их сравнение с аналогичными данными для эталонных текстов всех авторов. Для каждой пары «анализируемый текст – эталонный текст» (всего N пар по количеству авторов) введем числовую характеристику для оценки близости текстов. Автором анализируемого текста при таком подходе будет считаться тот, чей эталонный текст по заданной характеристике, т. е. расстоянию, окажется наиболее близок к анализируемому тексту. Такой подход предполагает выполнение гипотезы компактности: тексты одного автора образуют компактное множество в пространстве выбранных признаков [3; 4]. Предположение о выполнении данной гипотезы лежит в основе большинства алгоритмов распознавания [3; 4, с. 115]. Компактность понимается здесь не в математическом, а в «бытовом» смысле [2; 4, с. 115] и в данной задаче

означает, что гистограммы частот переходов двухбуквенных сочетаний для текстов одного автора должны быть похожи.

Выбор подходящего расстояния (меры схожести) в подобных задачах является достаточно сложной и мало изученной проблемой [2]. В [6] было проанализировано 5 способов подсчитать расстояние между текстами по используемым признакам (в т. ч. способ, предложенный Д. В. Хмелевым) и эмпирическим путем установлена наибольшая эффективность подсчета расстояния хи-квадрат [6; 9, p. 300, 303]:

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{k}\frac{\left(p_{ij} - q_{ij}\right)^2}{p_{ij} + q_{ij}}$$

$$p_{ij} = \frac{m_{1ij}}{n_1}; \quad q_{ij} = \frac{m_{2ij}}{n_2}$$

$$\sum_{i=1}^{k}\sum_{j=1}^{k} p_{ij} = \sum_{i=1}^{k}\sum_{j=1}^{k} q_{ij} = 1$$

m_{1ij} – число переходов из i-го двухбуквенного сочетания в j-ое в анализируемом тексте, n_1 – общее число переходов в анализируемом тексте, m_{2ij} и n_2 – аналогичные величины для эталонного текста, k – общее число двухбуквенных сочетаний в обоих текстах.

Так же предлагается использовать расстояние Топсе (Topsøe) из семейства энтропийных расстояний [9, p. 303; 12]:

$$tps = \sum_{i=1}^{k}\sum_{j=1}^{k} p_{ij}\log_2\left(\frac{2p_{ij}}{p_{ij} + q_{ij}}\right) + q_{ij}\log_2\left(\frac{2q_{ij}}{p_{ij} + q_{ij}}\right)$$

и расстояние Джеффриса-Матусита (Jeffries-Matusita), которое, хотя и принадлежит к тому же семейству «квадратичных» (L2) расстояний, что и хи-квадрат, но, в отличие от двух приведенных выше расстояний, является метрикой (в отличие от них для него выполнено неравенство треугольника) [9, p. 303; 12]:

$$jm = \left(\sum_{i=1}^{k}\sum_{j=1}^{k}\left(\sqrt{p_{ij}} - \sqrt{q_{ij}}\right)^2\right)^{\frac{1}{2}}$$

Автором текста будет считаться тот, у кого значение используемого расстояния для пары «анализируемый текст – эталонный текст» окажется наименьшим.

Предложенный алгоритм не зависит от языка, на котором написаны анализируемые тексты [5, с. 97] и, в силу вида используемых расстояний, имеет некоторую устойчивость к выбросам в данных, например, ошибкам оптического распознавания символов (OCR).

Верификация

Для верификации предлагаемой методики был подготовлен корпус патристических текстов на древнегреческом языке из 10 авторов со II по XII век. Использовались свободные распознанные тексты издания Patrologia Graeca (PG) [15] из Open Migne Project (тексты с ошибками распознавания и без вычитывания человеком). Ниже приводится список авторов и соответствующие им тексты, в скобках указывается номер текста по Clavis Patrum Graecorum (CPG):

1. Прав. Климент Александрийский: Protrepticus (CPG 1375) (частично), Stromata (CPG 1377) (частично); **2.** Свт. Афанасий Великий: Contra gentes (CPG 2090), De incarnatione Verbi (CPG 2091), Epistula ad episcopos Aegypti et Libyae (CPG 2092), In illud: Omnia mihi tradita sunt (CPG 2099), De decretis Nicaenae synodi (CPG 2120), De sententia Dionysii (CPG 2121), Apologia de fuga sua (CPG 2122), Apologia contra arianos sive apologia secunda (CPG 2123), Epistula encyclica (CPG 2124), Apologia ad Constantium imperatorem (CPG 2129); **3.** Свт. Епифаний Кипрский: Panarion (Adversus haereses) (CPG 3745) (частично); **4.** Свт. Василий Великий: Homiliae in Hexaemeron (CPG 2835), Homiliae super Psalmos (CPG 2836), Adversus Eunomium (CPG 2837) (первые 3 книги, частично), De Spiritu Sancto (CPG 2839); **5.** Свт. Григорий Нисский: Contra Eunomium (CPG 3135), Ad Ablabium quod non sint tres dei (CPG 3139), Contra fatum (CPG 3152), Homiliae in Ecclesiasten (CPG 3157); **6.** Свт. Григорий Богослов: Orationes (CPG 3010) (частично: слова 1-2, 3 (частично), 4-9, 14, 15 (частично), 16-19, 40-43); **7.** Свт. Иоанн Златоуст: Homiliae in Genesim (CPG 4409); **8.** Свт. Кирилл Александрийский: De adoratione et cultu in spiritu et veritate (CPG 5200) (частично); **9.** Преп. Иоанн Дамаскин: Dialectica (CPG 8041), Expositio accurata Fidei Orthodoxae (CPG 8043), Contra imaginum calumniatores orationes tres (CPG 8045); **10.** Свт. Феофилакт Болгарский: Enarratio in Evangelium Matthaei, Marci et Lucae.

Все тексты кодировались в UTF-8 в нормальной форме KD (normalization form KD – NFKD), т. к. в юникоде визуально неотличимые символы в контексте вычислительной среды могут считаться различными, а NFKD позволяет в этом смысле унифицировать тексты, полученные из разных источников (подробнее см. в [17]).

Для всех авторов их тексты объединялись через пробел в один текст. Во всех получившихся текстах удалялись все символы, не являющиеся буквами и пробелами (в т. ч. придыхания и ударения). Такой прием в т. ч. минимизирует предварительную ручную обработку входных текстов: например, пропадает необходимость удалять некоторые возможные ошибки автоматически распознанного текста в середине слова. Между словами оставлялось только по одному пробелу и все буквы в верхнем регистре переводились в нижний.

Тестирование точности описанной выше методики проводилось методом повторяющегося скользящего контроля по k-блокам (repeated k-fold cross-validation) [8, p. 331] – стандартной процедурой применяющейся для таких проверок и позволяющей при тестировании учесть всю имеющуюся выборку. Опишем ее применительно к нашим данным. Пусть есть 2 автора и соответствующие им тексты. Разобьем эти

тексты, например, на 3 последовательных равных по размеру фрагмента (количество фрагментов является одним из параметров метода тестирования и может меняться):

Автор 1:	1.1	1.2	1.3
Автор 2:	2.1	2.2	2.3

Для каждого автора каждый из получившихся фрагментов будем сравнивать с объединениями соответствующих оставшихся фрагментов. Для данного примера будут выполнены следующие сравнения: 1.1 с 1.2 ∪ 1.3, 1.1 с 2.2 ∪ 2.3 (т. е. при сравнении фрагмента 1.1 с другими фрагментами, соответствующий ему фрагмент 2.1 другого автора не учитывается; это позволяет достичь равенства эталонных фрагментов, с которыми происходит сравнение), 1.2 с 1.1 ∪ 1.3, 1.2 с 2.1 ∪ 2.3 и т. д. При разбиении всего текста на 3 фрагмента отношение анализируемого фрагмента и оставшихся объединений (эталонного текста) будет, соответственно, 1 к 2.

Для примера выше было рассмотрено последовательное разбиение текста. В повторяющемся скользящем контроле по k-блокам разбиение на равные фрагменты производится не последовательно, но начало нового фрагмента определяется псевдо-случайно [8, p. 331]. Если полученный индекс начала нового фрагмента указывал на пробел или на часть слова, то, по возможности, индекс передвигался в тексте вперед до начала нового слова для обеспечения более правдоподобного разбиения с целым словом (а не частью слова) в начале фрагмента.

Для каждого результата точности проводилось 1 000 запусков алгоритма. Точность определялась как процентное отношение правильных ответов ко всем ответам.

В столбцах таблиц 1 и 2 приводятся результаты точности (в процентах) работы описанного алгоритма в зависимости от величин анализируемого и эталонного текстов и использованного расстояния. Разбиение для каждого из авторов соответствующего им текста в 100 000 символов на 10 фрагментов означает, что анализируемый текст состоял из 10 000 символов, а эталонные тексты для каждого из авторов – из 90 000 (1 к 9).

Загрузка текстов для каждого автора происходила в той последовательности, как они перечислены выше при описании тестового корпуса текстов, и прекращалась при достижении необходимого количества символов (т. е. последний загруженный текст автора мог быть загружен лишь частично).

Общее кол.-во символов для текстов каждого автора	Разбиение на 2 фрагмента			Разбиение на 10 фрагментов		
	χ^2	tps	jm	χ^2	tps	jm
12 500	**92,9**	91,9	91,8	87,7	89,9	**91,1**
25 000	**92,1**	91,5	91,4	91,0	93,6	**94,3**
50 000	**95,5**	93,8	92,8	93,2	93,6	**94,8**
100 000	98,4	**98,6**	97,3	96,6	**97,4**	96,9
200 000	97,0	**97,7**	96,6	97,9	**98,7**	97,6
300 000	96,1	**96,5**	95,5	**97,3**	97,2	96,4
350 000	**98,5**	**98,5**	97,8	**98,5**	97,9	97,9
400 000	**99,5**	**99,5**	98,0	**99,0**	98,0	97,8
500 000	**99,8**	99,6	99,7	**98,7**	98,1	97,4
600 000	**99,1**	99,0	98,6	**98,9**	98,3	98,1
700 000	**99,2**	97,8	97,4	**98,4**	98,0	96,5

Таблица 1. Точность работы алгоритма в зависимости от величин анализируемого и эталонного текстов и использованного расстояния (разбиение на 2 и 10 фрагментов).

Общее кол.-во символов для текстов каждого автора	Разбиение на 20 фрагментов			Разбиение на 30 фрагментов		
	χ^2	tps	jm	χ^2	tps	jm
12 500	82,9	86,4	**88,8**	76,4	80,7	**85,4**
25 000	84,0	88,8	**91,8**	79,0	83,8	**90,5**
50 000	89,6	92,2	**93,7**	84,9	88,5	**92,0**
100 000	92,8	94,8	**95,8**	87,9	90,9	**94,8**
200 000	93,9	94,1	**94,7**	89,5	91,2	**92,3**
300 000	93,5	93,6	**93,8**	90,0	**90,8**	**90,8**
350 000	**95,8**	94,8	94,3	90,2	**91,2**	91,0
400 000	**96,5**	96,0	95,1	92,4	**92,5**	91,6
500 000	**97,6**	97,2	96,1	93,9	93,9	**94,4**
600 000	**97,9**	97,4	96,6	**96,0**	95,1	94,8
700 000	**97,1**	97,4	95,6	**95,6**	94,9	93,4

Таблица 2. Точность работы алгоритма в зависимости от величин анализируемого и эталонного текстов и использованного расстояния (разбиение на 20 и 30 фрагментов).

Из полученных результатов скользящего контроля можно сделать следующие выводы:

1) При использовании расстояния хи-квадрат и отношении анализируемого текста к эталонному как 1 к 9 (т. е. изначальный текст делился на 10 фрагментов) точность свыше 90% достигается с величины анализируемого текста в 2 500 символов, а эталонного – в 22 500. Меньшие величины исходных текстов при таком делении, видимо, не дают должного количества данных для более точного различения авторов. Такая точность обеспечивают возможность решать довольно широкий спектр потенциальных задач.

2) Метрика Джеффриса-Матусита показывает лучшие результаты при небольших объемах текстов и значительном различии величины анализируемого текста от эталонного (1 к 19 и 1 к 29), но при увеличении объемов текстов и сохранении их соотношений использование расстояния хи-квадрат дает большую точность.

3) Расстояние Топсе при небольших объемах текстов и значительном различии величины анализируемого текста от эталонного (1 к 19 и 1 к 29) оказалось лучше расстояния хи-квадрат, но хуже метрики Джеффриса-Матусита. В других тестах точность с расстоянием Топсе, в целом, довольно близка к результатам с хи-квадрат. Имея в виду, что расстояние Топсе принадлежит к другому семейству расстояний, чем хи-квадрат, первое можно использовать для перепроверки результатов последнего.

4) Для большинства результатов (кроме 17 из 132 – 12,9%), приведенных в таблицах 1 и 2, доля неправильных ответов, где истинный автор был на 2-м месте из 10 возможных, превышала 50%.

5) Хотя расстояние Джеффриса-Матусита является метрикой, его использование не везде дало лучшие результаты, хотя, теоретически, имея в виду, что в отличие от двух других рассмотренных расстояний для него выполняется неравенство треугольника, этого можно было бы ожидать.

Эффективность данного метода, похоже, связана с тем, что статистика по переходам двухбуквенных сочетаний содержит некоторую информацию о предпочитаемом автором стиле и синтаксических конструкциях, а также – об авторском словаре [5, с. 104-105].

При использованной процедуре скользящего контроля может случиться так, что анализируемый и эталонный тексты будут принадлежать одному произведению какого-либо автора. Такое деление при тестировании точности несколько искусственно и на практике довольно часто встречается задача анализа одного законченного произведения. Ниже проводится анализ авторства 35 отдельный текстов 3-х авторов, в качестве эталонов используется описанный выше корпус текстов из 10 авторов (по 700 000 символов для каждого). Анализируемые тексты (взяты из [19]):

1. Свт. Афанасий Великий: Vita s. Antonii (CPG 2101); **2.** Свт. Григорий Богослов: Orationes (CPG 3010) (частично: слова 10-13, 20-39, 44, 45); **3.** Свт. Иоанн Златоуст: Adversus judaeos orationes (CPG 4327) (8 слов).

При использовании расстояния хи-квадрат неверно было определено авторство 5 текстов из 35 анализируемых (14%), а именно: слова свт. Григория Богослова № 20, 29-31 и 35. Для всех этих неверно классифицированных слов автором был определен свт. Григорий Нисский. Для слов № 20 и 35 свт. Григорий Богослов оказался на втором месте претендентов на авторство (из 10 возможных), для слов № 29-31 на втором месте был преп. Иоанн Дамаскин. Для 21 верно определенного слова свт. Григория Богослова наиболее популярными авторами, оказавшимися на втором месте, были свт. Григорий Нисский (10 слов) и свт. Василий Великий (9 слов). Все слова с неправильно определенным авторством с расстоянием хи-квадрат так же были ошибочно классифицированы и с расстояниями Топсе и Джеффриса-Матусита. Кроме этого, с

последними двумя расстояниями было неправильно определено авторство слов свт. Григория Богослова № 22 и 28 как принадлежащих свт. Григорию Нисскому (в качестве возможных авторов свт. Григорий Богослов был определен на втором месте).

Определение авторства текстов De creatione hominis sermo 1, 2 (CPG 3215, 3216)

В качестве примера работы описанной методики для решения реальных задач атрибуции рассмотрим ее применение для определения авторства текстов De creatione hominis sermo 1, 2 (CPG 3215, 3216), которые в одних рукописях приписываются свт. Василию Великому, а в других – его младшему брату свт. Григорию Нисскому [1].

В таблице 3 приводятся данные об анализируемых и эталонных текстах для двух претендентов на авторство. Тексты Homiliae in Hexaemeron (CPG 2835) и De opificio hominis (CPG 3154) взяты из [10], а De creatione hominis sermo 1, 2 (CPG 3215, 3216) – из Open Migne Project.

Автор	Текст	№ CPG	Кол.-во символов
Эталонные тексты			
Свт. Василий Великий	Homiliae in Hexaemeron	2835	218 155
Свт. Григорий Нисский	De opificio hominis	3154	166 949
Анализируемые тексты			
	De creatione hominis sermo 1	3215	27 975
	De creatione hominis sermo 2	3216	25 436

Таблица 3. Данные об анализируемых и эталонных текстах.

В качестве эталонного текста для свт. Василия Великого были выбраны Homiliae in Hexaemeron (CPG 2835), поскольку анализируемые тексты (CPG 3215, 3216) предполагаются как их продолжение, а для свт. Григория Нисского – De opificio hominis (CPG 3154), так как это сочинение имеет близкую к анализируемым текстам тему. Представляется, что подбор эталонных текстов нужно стараться согласовывать с жанром и темой анализируемого текста (см.: [13, p. 580]). Например, используя в качестве эталона для одного автора произведение стихотворное, а для другого – написанное изречениями-сотницами, вполне можно получить нерелевантные результаты при определении авторства прозаической гомилии, поскольку такие эталонные тексты не будут давать репрезентативной статистической выборки для данной задачи, ведь их стиль заметно отличается не только между собой, но и с анализируемым текстом.

Перед анализом была проведена проверка эталонных текстов повторяющимся скользящим контролем по k-блокам с разбиением на 7 фрагментов, т. к. при таком значении размер выделяемых фрагментов текстов примерно соответствует размерам текстов анализируемых. Оказалось, что эталонные тексты хорошо различаются описанной методикой и точность проверки составила 100% для всех расстояний, т. е. на 1 000 запусках авторство всех фрагментов оказалось определено верно.

Для обоих текстов De creatione hominis sermo 1, 2 (CPG 3215, 3216) автором был определен свт. Василий Великий. Результаты вычисления всех расстояний для каждого автора приводятся в таблице 4.

Текст	№ CPG	Расстояния					
		Свт. Василий Великий			Свт. Григорий Нисский		
		χ^2	tps	jm	χ^2	tps	jm
De creatione hominis sermo 1	3215	0,7813	0,7200	0,8099	0,7883	0,7286	0,8174
De creatione hominis sermo 2	3216	0,7986	0,7403	0,8242	0,8307	0,7694	0,8401

Таблица 4. Результат работы алгоритма определения авторства текстов De creatione hominis sermo 1, 2 (CPG 3215, 3216). Автором считается тот, для кого вычисленное расстояние оказалось наименьшим. Для обоих текстов автором был определен свт. Василий Великий.

Библиография

1. Василий Великий / П.Б. Михайлов [и др.] // Православная энциклопедия. М.: Церковно-научный центр «Православная энциклопедия», 2004. Т. 7. С. 131-191.
2. Гипотеза компактности // MachineLearning.Ru. 2017. URL: http://www.machinelearning.ru/wiki/index.php?title=Гипотеза_компактности (дата обращения: 15.06.2017).
3. Гуров С.И., Потепалов Д.Н., Фатхутдинов И.Н. Решение задач распознавания с невыполненной гипотезой компактности // Математические методы распознавания образов (ММРО-13). – М.: МАКС Пресс, 2007. – С. 27-29.
4. Загоруйко Н.Г. Гипотезы компактности и λ-компактности в методах анализа данных // Сибирский журнал индустриальной математики. 1998. Т. 1. № 1. С. 114-126.
5. Кукушкина О.В., Поликарпов А.А., Хмелев Д.В. Определение авторства текста с использованием буквенной и грамматической информации // Проблемы передачи информации. 2001. № 37(2). С. 96-109.
6. Поддубный В.В., Шевелев О.Г., Фатыхов А.А. Сравнительный анализ эффективности алгоритмов распознавания авторства текстов по частотам переходов // Вестник Томского государственного университета. Серия “Математика. Кибернетика. Информатика”. 2006. № 290. С. 232-234.
7. Хмелев Д.В. Распознавание автора текста с использованием цепей А. А. Маркова // Вестник МГУ. Сер. 9: Филология. 2000. № 2. С. 115-126.
8. Alpaydin E. Introduction to machine learning. Cambridge, Massachusetts, London: MIT Press, 2004. 415 p.
9. Cha S.-H. Comprehensive survey on distance/similarity measures between probability density functions // International journal of mathematical models and methods in applied sciences. 2007. Vol. 1(4). P. 300-307.
10. Documenta catholica omnia. URL: http://www.documentacatholicaomnia.eu (24.07.2017).

11. Khmelev D.V., Tweedie F.J. Using Markov chains for identification of writers // Literary and linguistic computing. 2001. № 16(4). P. 299-307.
12. Kocher M., Savoy J. Distance measures in author profiling // Information processing & management. 2017. Vol. 53(5) [forthcoming]. P. 1103-1119.
13. Maspero G., Espositi M.D., Benedetto D. Who wrote Basil's epistula 38? A possible answer through quantitative analysis // Gregory of Nyssa: Contra Eunomium III. An english translation with commentary and supporting studies. 2014. P. 579-594.
14. N-gram-based author profiles for authorship attribution / V. Keselj [et al.] // Proceedings of the conference pacific association for computational linguistics (PACLING'03). 2003. P. 255-264.
15. Robertson B., Dalitz Ch., Schmitt F. Automated page layout simplification of Patrologia Graeca // Proceedings of the first international conference on digital access to textual cultural heritage (DATeCH'14). 2014. P. 167-172.
16. Stamatatos E. A survey of modern authorship attribution methods // Journal of the American society for information science and technology (JASIST). 2009. Vol. 60(3). P. 538-556.
17. Unicode 5.1.0 Standard Annex #15. Unicode normalization forms. 2008. URL: http://www.unicode.org/reports/tr15/tr15-29.html (дата обращения: 15.05.2017).
18. Who wrote the web? Revisiting influential author identification research applicable to information retrieval / M. Potthast [et al.] // Advances in information retrieval. 38th European conference on IR research (ECIR 2016). 2016. P. 393-407.
19. Βικιθήκη (Greek Wikisource). Κατηγορία: Συγγραφείς. URL: https://el.wikisource.org/wiki/Κατηγορία:Συγγραφείς (25.07.2017).