Настоящая работа посвящена проблеме численного оценивания близости тематического текста наиболее рациональному (эталонному) языковому варианту описания представляемого им фрагмента знаний. Данная проблема актуальна для реализации целенаправленного отбора текстовой информации без потери полезной смысловой составляющей. Примерами практических приложений здесь могут быть подбор статей для публикации в научных изданиях, а также разработка учебных курсов и образовательных порталов. Так, в процессе подготовки учебного материала преподаватель должен получить доступ к некоторому срезу информационного пространства, элементами которого являются публикации либо Internet-страницы, релевантные учебному курсу. Основное требование здесь можно представить как сортировку источников информации по степени отражения наиболее существенных понятий изучаемой предметной области при максимальной компактности и безызбыточности изложения (*плакат 2*). В идеале из информационных источников выстраивается иерархия, на верхний уровень которой выносятся те из них, с которых следует начинать изучение. Аналогичную возможность должен получить и студент в процессе самостоятельной работы, что особенно актуально в рамках обучения студентов подготовке и реализации проектов в профессиональной сфере.

В настоящей работе разбиением слов каждой фразы анализируемого текста на классы по значению меры TF-IDF решается задача оценки его близости наиболее рациональному (эталонному) языковому варианту передачи смысла без поиска перифраз. При этом в роли анализируемых текстов выступают аннотации научных статей вместе с их заголовками как отражающие основное содержание каждой из работ и наиболее значимые результаты без излишних методологических деталей.

Содержательно близкая задача (*плакат 3*) – построение и верификация иерархических тематических моделей крупных конференций и поиск наиболее релевантных тем для нового участника. При этом тема документа определяется его терминами из терминологического словаря конференции. Ключевым моментом здесь является значение важности термина, которое выражается через его энтропию относительно экспертной кластеризации на заданном уровне иерархии. Языковые выразительные средства, значимые для выбора лучшего варианта среди возможных перифраз, при таком подходе остаются вне рассмотрения. Известные решения в области детектирования перифраз и обучения такому детектированию также не предусматривают должного качественного анализа самих перифраз. В лучшем случае вычисляется степень смысловой близости предложений, например, с привлечением синтаксического анализатора SyntaxNet и определением редакционного расстояния между полученными деревьями зависимостей. Интерпретация природы перифраз, значимая для выбора наиболее рационального языкового варианта передачи фрагмента знаний, здесь не предусматривается. Однако даже качество подготовки корпуса перифраз с помощью системы, распознающей похожие по смыслу предложения, зависит от правильности её обучения. Первостепенную роль здесь играет выделение набора единиц текста и их связей, необходимого и достаточного для представления единицы знаний. Именно такой набор отвечает смысловому эталону.

В предлагаемом решении основой оценки близости текста эталону является разбиение слов каждой его фразы на классы по значению меры TF-IDF относительно текстов корпуса, предварительно формируемого экспертом (*плакаты 4–6*). Для отнесения сочетаний слов к ключевым из определяющих образ фразы в настоящей работе используется представленная на *плакате 4* интерпретация меры TF-IDF, оценивающей число одновременных вхождений всех слов анализируемого

сочетания во фразы отдельного документа корпуса (значение в числителе формулы (1)). При подсчете общего числа слов документа (знаменатель формулы (1)) здесь раздельно учитываются случаи совместной встречаемости слов сочетания и встречаемость без одновременного вхождения во фразу. При этом (*плакат 5*) значение TF-IDF ключевого сочетания слов должно быть не ниже минимального из значений указанной меры по его отдельным словам.

Используемый в работе вариант поиска необходимых и достаточных составляющих образа фразы предметно-ограниченного естественного языка в виде ключевых слов и их сочетаний, представленный на *плакате 7*, строится из следующих эмпирических соображений. Во-первых, разделение на общую лексику и термины здесь должно быть выражено как можно в большей степени, а слова в кластерах, формируемых по TF-IDF, должны быть распределены более или менее равномерно. Кроме того, число получившихся кластеров должно быть как можно ближе к трём при максимуме значений TF-IDF для слов кластера наибольших значений указанной меры. Данное требование следует понимать как максимальную релевантность терминов в составе фраз отбираемого документа сформированному корпусу. Сами документы корпуса сортируются по убыванию произведения представленных на *плакате 7* оценок, а в качестве оценки близости фразы эталону при этом берётся наибольшее из получившихся значений.

Для группы фраз, первая из которых – заголовок научной статьи, а остальные представляют аннотацию, в настоящей работе вводятся два варианта оценки близости эталону, в равной мере предусматривающие минимум среднеквадратического отклонения (СКО) значения близости эталону по всем фразам группы.

Первый вариант (*плакат 8*) подразумевает максимальную близость эталону для заголовка статьи. Отметим, что введённая оценка не подразумевает сортировку фраз группы по близости эталону и содержательно соответствует порядку отбора статей, начиная с анализа заголовка. Такая постановка задачи наиболее адекватна общепринятому в научной периодике требованию отражения в заголовке содержания статьи. Однако априорное предположение о максимальной близости эталону именно заголовка статьи на практике выполняется не всегда.

Учитывая вышесказанное, *второй вариант* (*плакат 9*) использует в числителе расчётной формулы максимальное из полученных значений оценки по всем фразам анализируемого текста. При этом максимальный итоговый рейтинг по коллекции получает статья с наибольшим значением *первого варианта* оценки, попадающим в один кластер со значением *второго варианта* оценки для той же статьи. Корректное применение данного утверждения предполагает отнесение к одному кластеру значений *первого* варианта оценки для статьи, получившей максимальный итоговый рейтинг, и максимального значения *первого варианта* оценки по коллекции, из которой производится отбор. В случае отсутствия в коллекции статьи, отвечающей данному требованию, максимальный итоговый рейтинг получает статья с наибольшим значением *первого варианта* оценки по анализируемой коллекции.

Как видно из определения, оба варианта оценки зависят от подбора корпуса экспертом. Кроме того, поскольку заголовок и фразы аннотации статьи (по определению) несут некий единый смысловой образ, то вполне допустима мена местами рассматриваемых вариантов оценки в *Утверждении 2* на *плакате 9*.

Экспериментальный материал для апробации метода приведён на *плакатах 10–12*. Программная реализация на языке Python 2.7 и результаты экспериментов представлены на портале Новгородского университета. Основным критерием при

выборе коллекций, как и при подборе текстов в корпус, была максимально полная и наглядная иллюстрация разделения слов на общую лексику и термины. В целях более точного выделения смыслового контекста терминов вычисление меры TF-IDF слов анализируемых фраз производилось без учёта предлогов и союзов.

Представленные далее в таблицах на *плакатах 13–17* результаты экспериментов являются подтверждением правила «хорошего тона» ряда изданий по информатике и вычислительной технике отображать в заголовке название метода, модели, алгоритма, представляемого в работе, а также теоретическую основу предлагаемых решений. По коллекции *«ММРО-15, Статистическая теория обучения»* максимумы обоих вариантов оценки близости эталону были получены для одной и той же статьи, как и по коллекции *«ММРО-15, Математическая теория и методы классификации»*. Как видно из таблиц на *плакатах 13–16*, значения вариантов оценок близости эталону у указанных статей совпадают. Таким образом, согласно условию *Утверждения 2*, статьи *К.В. Воронцова, Г.А. Махиной «Принцип максимизации зазора для монотонного классификатора ближайшего соседа»* и *И.Е. Генрихова, Е.В. Дюковой «Полные решающие деревья в задачах классификации по прецедентам»* получают максимальный рейтинг каждая по своей коллекции.

Результат, полученный по коллекции *«ММРО-14, Методы и модели распознавания и прогнозирования»* и представленный на *плакате 16*, иллюстрирует случай, когда статья с наибольшим значением *второго варианта* оценки близости эталону по коллекции имеет не попадающее с ним в один кластер значение *первого варианта* данной оценки. Действительно, для статьи *Д.И. Мельникова, В.В. Стрижова, Е.Ю. Андреевой, Г. Эденхартера «Выбор опорного множества при построении устойчивых интегральных индикаторов»* значения *первого* и *второго вариантов* оценки, равные, соответственно, 0.0129 и 0.1426, образуют два самостоятельных кластера. В силу сказанного максимальный итоговый рейтинг по коллекции будет у статьи *О.В. Бариновой и Д.П. Ветрова*, получившей максимальное по коллекции значение *первого варианта* оценки близости эталону (*плакат 17*).

Аналогичная ситуация имеет место и по коллекции *«ИОИ-9, Математическая теория и методы классификации»*. Здесь максимальное значение *второго варианта* оценки, равное 0.1336, получила статья *Н.К. Животовского и К.В. Воронцова «Критерии точности комбинаторных оценок обобщающей способности»*. Значение *первого варианта* оценки, равное 0.0600, здесь попадает в один кластер с максимальным значением данной оценки по коллекции, равным 0.0920, но, тем не менее, не лежит в одном кластере со значением *второго варианта* оценки близости эталону для данной статьи. Поэтому максимальный итоговый рейтинг получает статья *С.Д. Двоенко и Д.О. Пшеничного* с наибольшим значением *первого варианта* оценки по рассматриваемой коллекции.

В случае мены местами вариантов оценки близости эталону в *Утверждении 2* на *плакате 9* по коллекциям *«ММРО-15, Статистическая теория обучения»* и *«ММРО-15, Математическая теория и методы классификации»* при этом максимальные рейтинги получат те же самые статьи (*плакат 17*). По коллекции *«ММРО-14, Методы и модели распознавания и прогнозирования»* максимальный рейтинг здесь снова получит статья *О.В. Бариновой и Д.П. Ветрова*. Действительно, максимальное по коллекции значение *второго варианта* оценки близости эталону будет у статьи *Д.И. Мельникова, В.В. Стрижова, Е.Ю. Андреевой, Г. Эденхартера*. Но как было показано нами ранее, значения *первого* и *второго вариантов* оценки близости эталону для данной статьи в рассматриваемой коллекции относят-

ся к разным кластерам. Следовательно, согласно условию *Утверждения 2* максимальный итоговый рейтинг по коллекции здесь получит статья, имеющая из оставшихся статей максимальное значение *второго варианта* оценки, попадающее в один кластер со значением *первого варианта* оценки близости эталону для неё же самой, то есть статья *О.В. Бариновой и Д.П. Ветрова*.

Единственным исключением в рассматриваемой серии экспериментов будет результат по коллекции *«ИОИ-9, Математическая теория и методы классификации»*. Как и в предыдущем примере, максимальный рейтинг по коллекции могла бы получить статья с максимальным значением *второго варианта* оценки, лежащим в одном кластере со значением *первого варианта* данной оценки, то есть работа *С.Д. Двоенко и Д.О. Пшеничного*. Но значение *второго варианта* оценки по этой статье не принадлежит тому же кластеру, что и максимальное по коллекции значение данного варианта оценки, которое вместе с максимальным итоговым рейтингом здесь получает статья *Н.К. Животовского и К.В. Воронцова*.

Отметим, что полученные результаты подтверждают гипотезу относительно смысловой нагрузки заголовка научной публикации по информатике и вычислительной технике. Для спорных случаев, подобных рассмотренному выше для коллекции *«ИОИ-9, Математическая теория и методы классификации»*, в зависимости от предметной области анализируемых текстов можно отдать предпочтение требованию отнесения к кластеру максимального значения либо *первого*, либо *второго варианта* оценки близости эталону.

Смысловой образ статьи с максимальным значением используемой оценки близости эталону по коллекции, из которой производится отбор, содержательно будут определять слова кластера наибольших значений TF-IDF относительно документа с наибольшим значением произведения представленных на *плакате 7* оценок, расположенные по соседству в линейном ряду соответствующей фразы анализируемой группы. В целях более точной идентификации многословных терминов на фоне общей лексики расширим выделяемые во фразе ключевые сочетания словами «серединного» кластера последовательности, формируемой на основе TF-IDF слов анализируемой фразы относительно заданного документа (*плакаты 18–19*).

Для подтверждения наличия в анализируемых фразах связей слов из кластеров наибольших значений TF-IDF был задействован MaltParser – инструмент для синтаксического разбора фраз естественных языков и работы с деревьями зависимостей. Как видно из таблиц на *плакатах 18–22*, расположению слов указанных кластеров по соседству в линейном ряду фразы соответствует синтаксическая связь, что говорит и о единстве составляющей смыслового образа текста.

Результат, полученный по коллекции *«ММРО-15, Математическая теория и методы классификации»*, полностью согласуется с теоретическим выводом об относительности понятия «общая лексика». Фактически каждая из представленных в таблицах коллекций относится к определённой теме, описываемой дискретным распределением на множестве терминов. Классификация же по TF-IDF здесь относит слова *«обзор»* и *«дать»*, рассматриваемые как общая лексика по языку в целом, но уникальные для статьи *И.Е. Генрихова* и *Е.В. Дюковой*, к терминам. Отметим, что сочетание указанных слов не отвечает условию *Утверждения 1* на *плакате 5* и, следовательно, не может быть отнесено к ключевым.

Поскольку заголовок и фразы аннотации отвечают единому смысловому образу, то вполне корректно анализировать вхождение слов, отнесённых к кластеру наибольших значений TF-IDF по одной фразе, в связи слов относительно других

фраз. В настоящей работе для отождествления набора таких связей с ключевым сочетанием слов вводится следующее условие: совокупности рассматриваемых сочетаний должен соответствовать связный подграф дерева синтаксического разбора фразы и минимум одно сочетание должно отвечать условию *Утверждения 1* на *плакате 5*. В примере для вышеупомянутой статьи *И.Е. Генрихова* и *Е.В. Дюковой* слово *«полный»* не вошло в кластер наибольших значений TF-IDF, но синтаксически подчинено слову *«дерево»* из указанного кластера и, следовательно, образует искомое ключевое сочетание. Следует отметить, что последний пример на *плакате 21* учитывает транзитивность синтаксического отношения в рамках последовательности соподчиненных слов, ср. «*матриц – сравнений – парных*». Отнесение рассматриваемого сочетания слов к ключевым на основе введенной интерпретации меры TF-IDF служит дополнительным подтверждением наличия связи. Отдельного исследования здесь заслуживает динамика изменения меры TF-IDF при переходе от дискретных слов к *L*-граммам (по К. Шеннону). Второй пример сверху на том же плакате является иллюстрацией того, что расположение слов кластера наибольших значений TF-IDF по соседству в линейном ряду фразы при этом следует рассматривать как необходимое, но не достаточное условие отнесения к ключевым сочетаниям, определяющим смысловой образ текста. Достаточные же условия здесь определяются предложенной в работе методикой использования введенной авторами интерпретации меры TF-IDF.

Основной результат настоящей работы – *метод оценки близости текста смысловому эталону относительно тематического текстового корпуса*.

Эффективность предложенного метода может быть оценена разбиением текстов коллекции на кластеры по значению оценки близости эталону для группы фраз и отношением числа текстов, отнесённых к кластеру наибольших значений данной оценки, к общему числу текстов в составе коллекции. Так, на материале коллекций, упомянутых на *плакатах 13–22,* имеем минимум трёхкратное сокращение числа документов (научных статей), с которыми следует ознакомиться в первую очередь при изучении заданной предметной области, например, студентами.

Учитывая оцениваемую при отнесении фразы к «эталонной» степень разделения её слов на общую лексику и термины, представляет также интерес интеграция оценок близости эталону для групп фраз по всей анализируемой коллекции статей. Присутствие ключевых сочетаний слов в аннотациях и заголовках при этом может служить основой при назначении итогового рейтинга статьи в спорных случаях, а также иерархизации самих статей по степени значимости для изучения заданной предметной области.