Настоящая работа посвящена проблеме численного оценивания взаимной смысловой зависимости тематических текстов относительно наиболее рациональных (эталонных) вариантов описания представляемых ими фрагментов знаний. Данная проблема актуальна при определении значимости источников информации относительно решаемых пользователем задач. При этом сортировка источников по степени отражения наиболее существенных понятий заданной предметной области при максимальной компактности и безызбыточности изложения предполагает построение иерархии, на верхний уровень которой выносятся те из источников, с которых следует начинать рассмотрение. Основой построения иерархии документов при этом будет взаимосвязь их смысловых эталонов таким образом, что эталон вышестоящего документа должен доопределять эталон непосредственно связанного с ним нижестоящего. Данное требование особенно актуально при формировании индивидуальной образовательной траектории обучаемого в электронном обучении.

Идейно близкая иерархия первоисточников естественным путём возникает при описании предметной области посредством тезауруса либо онтологии, поскольку их построение подразумевает интеграцию и систематизацию существующих источников информации по заданной тематике (*плакат 3*, [ВЦ РАН, тезаурус «Чёрный квадрат»]). Наиболее актуальная при таком подходе задача отмечена К. В. Воронцовым как выстраивание рекомендуемого порядка работы с источниками, включая поиск «точки входа». Для классификации документов здесь могут использоваться разные критерии, например, распределение документа по темам [Стрижов В. В., 2014], сравнение частоты встречаемости терминов в анализируемом документе и заданном референтном корпусе [Еремеев М. А., 2015] и многие другие. Упорядочивание источников от простого к сложному подразумевает анализ распределений частот встречаемости как отдельных слов, так и их сочетаний. Причём помимо отражения наиболее значимых понятий, минимума терминов с аномально высокой частотой по сравнению с референтным корпусом, существенную роль играют языковые выразительные средства, определяющие лучший вариант среди возможных перифраз текста. Содержательно здесь требуется выделить и проанализировать набор текстовых единиц и их связей, необходимый и достаточный для представления единицы знаний и отвечающий смысловому эталону.

В настоящей работе рассматривается решение данной задачи на основе совместного использования разбиения слов на классы по значению меры TF-IDF и оценки силы связи слов в рамках фраз анализируемых текстовых документов.

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 клю-

чевых 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 содержательно есть заголовок научной статьи, а остальные представляют аннотацию, в настоящей работе используются два ранее предложенных авторами 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*.

Смысловой образ статьи с максимальным значением используемой оценки близости эталону по коллекции, из которой производится отбор, содержательно определяют 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 выделяемые во фразе ключевые сочетания расширяются словами «серединного» кластера последовательности, формируемой на основе TF-IDF слов анализируемой фразы относительно заданного документа.

Альтернативой и дополнением меры TF-IDF при оценивании значимости сочетания слов в данной работе служит предложенная авторами ранее оценка значимости *n*-граммы для ранжирования документов. В качестве оценки «силы» связи слов в настоящей работе берётся представленная на *плакате 10* *оценка (8)*, содержательно близкая коэффициенту Танимото. Из оценок силы связи слов в дистрибутивно-статистическом методе построения тезаурусов данная оценка наиболее наглядна, но в то же время учитывает встречаемость каждого слова в отдельности.

Порядок выделения *n*-граммы на последовательности пар слов исходной фразы представлен *Определением 1* на *плакате 10*. При этом оценка силы связи для пары слов исходной фразы вычисляется только в том случае, если значение TF-IDF минимум одного из слов пары принадлежит либо первому, либо «серединному» кластеру, что в целом согласуется с выделением ключевых сочетаний на основе TF-IDF. Значимость сочетания слов, рассматриваемого как *n*-грамма (*формула (9)* на *плакате 11*) оценивается из геометрических соображений и подразумевает максимизацию суммы силы связи слов в его составе при минимуме среднеквадратического отклонения указанной величины по всем связям слов в составе *n*-граммы. Сама оценка вычисляется для *n*-грамм с ненулевым значением меры TF-IDF относительно документа с наибольшим значением произведения представленных на *плакате 7* оценок по заданному корпусу. Отнесение сочетания слов к ключевым на основе оценки значимости *n*-граммы строится из следующих эмпирических соображений (*плакат 11*). Множество найденных сочетаний разбивается на кластеры по значению указанной оценки, а в качестве ключевых берутся сочетания, не попадающие в кластер её наименьших значений.

Совместное использование *введенной интерпретации меры TF-IDF для сочетаний слов* и их классификации на основе оценки силы связи заключается в следующем (*плакат 11*): ключевым будет считаться сочетание слов, если оно либо отвечает условию относительно TF-IDF (*Утверждение 1 на плакате 5*), либо не попадает в кластер наименьших значений оценки значимости для *n*-грамм.

Экспериментальный материал для апробации метода приведён на *плакатах 12–14*. 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.

Из представленных на *плакате 14* коллекций для отбора статей далее в таблицах на *плакатах 15–18* представлены результаты экспериментов по коллекции для раздела «Статистическая теория обучения» сборника трудов 15-й Всероссийской конференции «Математические методы распознавания образов» (ММРО, 2011 г.). Отметим, что в *Таблице 2* на *плакате 15* представлены те документы данной коллекции, относительно которых минимум по одному сочетанию слов была вычислена оценка значимости сочетания слов как *n*-граммы (*формула (9)* на *плакате 11*). В описываемых экспериментах по документу с порядковым номером $i = 1$ из *Таблицы 1* на *плакате 15* максимум произведения представленных на *плакате 7* оценок при выделении ключевых сочетаний слов из текстов исходной коллекции достигался наибольшее число раз. В целях более наглядной иллюстрации разделения сочетаний слов на кластеры по значению оценки значимости *n*-граммы в *Таблице 5* на *плакате 18* приведены результаты кластеризации относительно именно

этого документа, исходная текстовая коллекция для выделения сочетаний здесь дополнена документами раздела «Математическая теория и методы классификации» вышеупомянутой конференции ММРО-15. Строки в *Таблице 3* на *плакате 16* и в *Таблице 5* на *плакате 18*, для сочетаний слов, не отвечающих условию отнесения к ключевым на основе значения TF-IDF (*Утверждение 1 на плакате 5*), выделены более тёмным фоном.

Примером сочетания слов, признанного экспертом как верное, но не отвечающего условию отнесения к ключевым на основе TF-IDF, может послужить сочетание №6 из *Таблицы 3* на *плакате 16*. Отметим (см. *плакат 17*), что данное сочетание попадает в кластер наибольших значений оценки значимости для *n*-грамм относительно документа с порядковым номером $i = 2$ по *Таблице 1* на *плакате 15* и, следовательно, может рассматриваться как ключевое.

Сочетание №10 из *Таблицы 3* на *плакате 16* не признано экспертом как значимое для передачи отождествляемой с исходной фразой единицы знаний, не отвечает условию отнесения к ключевым на основе TF-IDF и попадает (*плакат 17*) в кластер наименьших значений оценки значимости для *n*-грамм относительно документа с порядковым номером $i = 1$ по *Таблице 1* на *плакате 15*. В *Таблице 5* на *плакате 18* то же сочетание также отнесено к кластеру наименьших значений оценки значимости для *n*-грамм, но в этот же кластер вместе с ним попадают сочетания *«обучающий материал»* и *«расширение семейства»*, допускаемые экспертом.

Сочетания №15 из *Таблицы 3* и №18 из *Таблицы 5* признаны экспертом значимыми для передачи соответствующих единиц знаний, но попадают (*плакат 17* и *18*) в кластеры наименьших значений оценки значимости для *n*-грамм каждое относительно своего документа. Но, тем не менее, оба сочетания отвечают условию отнесения на основе TF-IDF и, следовательно, могут отождествляться с ключевыми.

Как видно из представленных на *плакатах 16–18* результатов, применение классификатора на основе оценки значимости для *n*-грамм как дополнения классификации на основе TF-IDF в значительной мере зависит от документа, относительно которого вычисляется оценка значимости *n*-граммы. Для повышения точности выделения ключевых сочетаний слов здесь требуется дополнительная статистика по документам, относительно которых достигался максимум близости эталону по различным фразам разных документов анализируемой коллекции. Более точное выделение ключевых сочетаний слов на основе меры TF-IDF обусловлено вычислением IDF-меры для сочетания относительно всего заданного текстового корпуса. По той же причине выделение рассматриваемых сочетаний на основе TF-IDF представляется более целесообразным с позиции основной задачи – поиска наиболее значимых составляющих образа исходной фразы. В целях более точного разделения слов на общую лексику и термины представляет также интерес исследование связи распределений частот встречаемости слов в кластерах наибольших значений TF-IDF по фразам разных текстов анализируемой коллекции и случаев достижения максимума произведения представленных на *плакате 7* оценок относительно конкретных документов заданного текстового корпуса.