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 решение данной задачи строится на базе предложенных авторами ранее оценок близости тематического текста смысловому эталону и классификации слов фраз сравниваемых текстов по значению меры TF-IDF как основы указанных оценок. Анализируемыми текстами являются аннотации научных статей вместе с их заголовками, а сам текст не перефразируется с целью поиска всех возможных семантически эквивалентных языковых форм описания единицы знаний.

Базовые идеи и предположения классификации слов исходной фразы по TF-IDF как основы оценки её близости смысловому эталону представлены на *плакатах 4–6*. Для отнесения сочетаний слов к ключевым из определяющих образ фразы в настоящей работе используется представленная на *плакате 4* интерпретация меры TF-IDF, оценивающая число одновременных вхождений всех слов анализируемого сочетания во фразы отдельного документа тематического 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. Само значение TF-IDF ключевого сочетания слов (*плакат 5*) 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 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.

Принимая во внимание указанные теоретические выводы, ранжирование статей на основе совместного использования обоих вариантов оценки близости эталону можно формально представить алгоритмом, приведённым на *плакате 10*. Для построения иерархии документов коллекции на выходе алгоритма используется аналогия с задачей вероятностного тематического моделирования, где иерархия тем моделирует стратегию поиска с постепенным фокусированием внимания пользователя на подтемах. Применительно к значениям меры TF-IDF ключевых терминов в нашей задаче это выражается в более высоких значениях TF-IDF этих терминов в родительском документе по сравнению с дочерним в формируемой иерархии.

Пусть $\mathbf{Ts}_i$ и $\mathbf{Ts}_j$ – тексты из входящих в отсортированную коллекцию на выходе алгоритма на *плакате 10*, причём $i > j$, то есть итоговый рейтинг статьи,

отвечающей группе фраз $\mathbf{Ts}_i$, выше аналогичного показателя для $\mathbf{Ts}_j$. Тогда дополняемость текста $\mathbf{Ts}_j$ текстом $\mathbf{Ts}_i$ относительно их смысловых эталонов определяется (*плакат 11*) долей слов кластеров наибольших значений TF-IDF фраз текста $\mathbf{Ts}_i$, не входящих в кластеры наибольших значений указанной меры по фразам текста $\mathbf{Ts}_j$, но, тем не менее, имеющих относительно тех же фраз ненулевые значения TF-IDF. Данное предположение естественным образом согласуется с известной в лингвистике дистрибутивной гипотезой, согласно которой смысл слова определяется его контекстом, то есть тем, в окружении каких слов оно появляется в большом корпусе текстов. В нашем случае в роли контекста выступает множество всех слов кластеров наибольших значений TF-IDF всех фраз текста $\mathbf{Ts}_i$.

Смысловые образы наиболее близких эталону текстов определяют слова с наибольшими значениями TF-IDF, которые при расположении по соседству в линейном ряду фразы с наибольшей вероятностью связаны по смыслу и образуют ключевые сочетания вместе со словами, близкими среднему значению указанной меры. При оценке дополняемости текста $\mathbf{Ts}_j$ текстом $\mathbf{Ts}_i$ рассматриваются (*плакат 12*) ключевые сочетания, найденные для текста $\mathbf{Ts}_i$ и включающие слова, с ненулевыми значениями TF-IDF, не входящие в кластеры наибольших значений указанной меры по фразам текста $\mathbf{Ts}_j$. При этом для каждого такого сочетания минимум одно слово должно принадлежать кластеру наибольших значений минимум для одной фразы текста $\mathbf{Ts}_i$.

Для подтверждения наличия связи смыслового эталона текста $\mathbf{Ts}_j$ с эталоном текста $\mathbf{Ts}_i$ в дополнение к оценкам (9) и (10) в работе используется приведённый на *плакате 13* вариант оценки представленности слов анализируемой фразы в первом, последнем и «серединном» кластерах последовательности, сформированной на основе TF-IDF её слов. Данная оценка (формула 11) строится из геометрических соображений, а именно: посредством деления числа слов в кластере на длину фразы здесь формализовано требование максимальной концентрации слов анализируемой фразы в трёх наиболее значимых для оценки её близости эталону кластерах. При этом для учёта связи текста $\mathbf{Ts}_j$ с текстом $\mathbf{Ts}_i$ по каждой его фразе кластер слов наибольших значений TF-IDF дополняется словами аналогичных кластеров фраз текста $\mathbf{Ts}_i$, не входящих в кластер наибольших значений указанной меры по анализируемой фразе, но имеющих по той же фразе ненулевые значения TF-IDF. Из последнего же и «серединного» кластеров по анализируемой фразе указанные слова удаляются (при их наличии там). Модифицированный указанным образом вариант оценки (11), представляемый формулой (12) на *плакате 13*, используется далее на *плакате 14* для формализации критерия выбора вышестоящего текста $\mathbf{Ts}_i$ для заданного текста $\mathbf{Ts}_j$ в формируемой иерархии документов. Непосредственно степень дополнения эталона текста $\mathbf{Ts}_j$ определяется сравнением значений представленных на *плакате 14* оценок (15) и (16) с соответствующими им оценками (13) и (14) согласно *Утверждению 3*. Данные оценки идейно близки представленным на *плакатах 8* и *9* оценкам близости эталону для группы фраз.

Экспериментальный материал для апробации метода приведён на *плакатах 15–17*. 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 15* коллекций для отбора статей далее на *плакатах 18–26* приводятся результаты экспериментов по коллекции для раздела «Статистическая теория обучения» сборника трудов 15-й Всероссийской конференции «Математические методы распознавания образов» (ММРО, 2011 г.). На *плакате 18* представлен результат ранжирования коллекции согласно алгоритму на *плакате 10* относительно *первого* варианта оценки близости эталону (*плакат 8*). В целях более наглядной иллюстрации работы вышеуказанного алгоритма на *плакате 19* приведён результат ранжирования той же коллекции, но относительно *второго* варианта оценки близости эталону (*плакат 9*). Жёлтым фоном в таблицах на *плакатах 18* и *19* выделены соответствующие графы для документов с несовпадающими значениями первого и второго варианта оценки близости эталону. Связи документов внутри иерархии в последующих таблицах и на рисунках в направлении от нижестоящего к вышестоящему интерпретируются как $j \to i$, где i и j – порядковые номера документов по таблице на *плакате 18*. В случае пустого множества ключевых сочетаний у вышестоящего документа связь исключается из рассмотрения (*плакат 22*) при нулевом значении оценки дополняемости текста $\mathbf{Ts}_j$ текстом $\mathbf{Ts}_i$, представленной на *плакате 11*. Связь не рассматривается также при одновременно нулевых значениях оценок дополняемости как с учётом, так и без учёта ключевых сочетаний слов. В *таблицах 3*, *5* и *6* графы для связей, отвечающих условию *Утверждения 3*, выделены зелёным цветом, для частично отвечающих данному условию – жёлтым. Полученные результаты в целом соответствуют выдвинутой нами гипотезе о характере перераспределения слов в кластерах, лежащей в основе оценки (12) на *плакате 13*. Один из вариантов более точного анализа здесь может быть основан на использовании квантилей эмпирических распределений частот встречаемости слов в первом, последнем и «серединном» кластерах по TF-IDF по разным текстам анализируемой коллекции. Сказанное востребовано при анализе помимо заголовка и аннотации, например, обзорной части научной статьи.

Примером использования *Утверждения 3* для выбора из нескольких вариантов вышестоящего текста для заданного в формируемой иерархии могут послужить связи $8 \to 4$ и $8 \to 7$. В рассматриваемой коллекции это есть выбор вышестоящего документа в иерархии для статьи *Ботова П.В. (8)* из двух вариантов: статья *Фрея А.И. (4)* и *Неделько В.М. (7)*. Для связи $8 \to 4$ значение оценки (9) на *плакате 11* равно 0.4, оценка (10) на *плакате 12* принимает значение 0.(3). По связи $8 \to 7$ имеем значение оценки (9), равное 0.(3), оценка (10) для данной связи не вычисляется в силу того, что для вышестоящего текста ключевые сочетания слов не найдены, фактически оценка (10) по определению совпала бы здесь с оценкой (9). Таким образом, при прочих равных условиях неубывания оценок (15) и (16) по отношению к соответствующим им оценкам (13) и (14) на *плакате 14* предпочтение отдаётся связи $8 \to 4$. Заметим также, что для указанной связи в число слов кластеров наибольших значений TF-IDF для фраз вышестоящего текста, не входящих в аналогичные кластеры по нижестоящему, но имеющих относительно его же фраз ненулевые значения указанной меры, войдут слова *“минимизация”* и *“риск”*, а для

связи $8 \to 7$ – только *“риск”*, что также служит дополнительным подтверждением вышеупомянутой гипотезы о характере перераспределения слов в кластерах.

Основной результат данной работы следует обозначить как *методику иерархизации текстов предметно-ограниченного естественного языка на основе оценок близости тематического текста смысловому эталону*.

Эффективность предложенного решения может быть оценена по числу и виду компонент связности графа, полученного из графа найденных связей между документами анализируемой коллекции путём замены ориентированных рёбер неориентированными. В идеале каждая компонента связности при восстановлении ориентации рёбер будет направленным деревом, где для любой вершины максимальная и минимальная высота дочернего поддерева различаются не более чем на 1 (по аналогии со сбалансированным по высоте двоичным деревом поиска), а число самих компонент связности должно быть как можно ближе к 1. После удаления из исходного графа тех связей, которые либо не отвечают условию *Утверждения 3*, либо по которым для вышестоящего текста ключевые сочетания слов не найдены (*плакат 23* и *25*), подграф, отвечающий максимальной компоненте связности, среди вершин с максимальной степенью всегда содержит вершину для статьи с максимальным итоговым рейтингом по коллекции. Для сравнения: вариант без учёта ключевых сочетаний слов по рассматриваемой коллекции из 10 статей даёт максимальную компоненту связности из 8 вершин, в случае учёта ключевых сочетаний – из 6 вершин. При этом за счёт статей, не нашедших отражения в максимальной компоненте связности, имеем дополнительное (минимум на 20%) сокращение числа документов, с которыми следует ознакомиться в первую очередь при изучении заданной предметной области, например, студентами.

В целях повышения точности разделения слов на общую лексику и термины представляет интерес *исследование связи* распределений частот встречаемости слов в кластерах наибольших значений меры TF-IDF по фразам разных текстов анализируемой коллекции и случаев достижения максимальной близости фраз эталону относительно конкретных документов заданного корпуса по значению произведения представленных на *плакате 7* оценок.