8. ЛЕКЦИЯ. Тематическое моделирование

8.1 Векторная модель текста.

Большинство современных алгоритмов индексации и поиска в той или иной степени основано на векторной модели текста, предложенной Дж. Солтоном в 1973 году. В векторной модели каждому документу приписывается список терминов, наиболее адекватно отражающих его смысл. Иными словами, каждому документу соответствует вектор, размерность которого равна числу терминов, которыми можно воспользоваться при поиске.

Для дальнейшего изложения введем несколько важных понятий: словарь, поисковый образ документа, информационный массив.

Словарь – это упорядоченное множество терминов. Мощность словаря обозначается как D.

Поисковый образ документа – это вектор размерности D. Самый простой поисковый образ документа – двоичный вектор. Если термин входит в документ, то в соответствующем разряде этого двоичного вектора проставляется 1, в противном же случае – 0. Более сложные поисковые образы документов связаны с понятием относительного веса терминов или частоты встречаемости терминов.

Любой запрос также является текстом, а значит, его тоже можно представить в виде вектора q. В процессе работы поискового алгоритма происходит сравнение векторов поискового образа документа и поискового образа запроса. Чем ближе вектор документа находится к вектору запроса, тем более релевантным он является. Обычно все операции информационного поиска выполняются над поисковыми образами, но при этом их, как правило, называют просто документами и запросами.

Информационный массив называют также информационным потоком, набором документов или коллекцией документов. Описанная модель информационного массива является наиболее широко используемой. В первую очередь это связано с простотой реализации и, как следствие, возможностью быстрой обработки больших объемов документов.

Матрица информационного массива изображена на рис. 1, где W_{ij} – вес термина t_j в документе d_i.

Для полного определения векторной модели необходимо указать, каким именно образом будет отыскиваться вес терма в документе. Существует несколько стандартных способов задания функции взвешивания:

- булевский вес – равен 1, если терм встречается в документе и 0 в противном случае;

- TF (term frequency, частота терма) — вес определяется как функция от количества вхождений терма в документе;
- TF-IDF (term frequency — inverse document frequency, частота терма — обратная частота документа) — вес определяется как произведение функции от количества вхождений терма в документ и функции от величины, обратной количеству документов коллекции, в которых встречается этот терм.

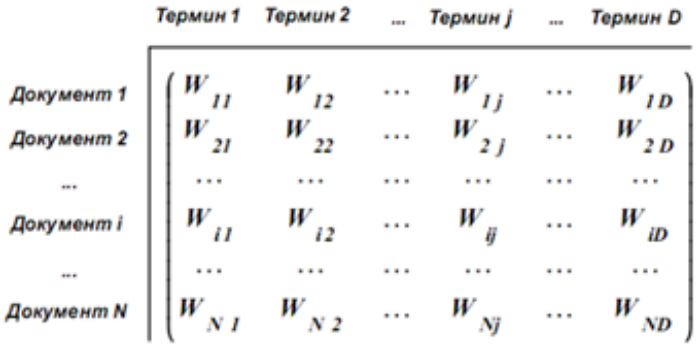

Рис. 1. Матрица «термин-документ» информационного массива

Остановимся подробнее на статистических закономерностях, которые используются в процессе индексирования документов.

8.2. Статистический анализ текстов. Закон Ципфа

Начальным этапом любого метода индексирования является отбор из документов терминов, которые бы наилучшим образом характеризовали их содержимое. Такая необходимость вызвана тем, что непосредственное сканирование текстов документов во время поиска занимает слишком много времени, особенно в поисковых системах сети Интернет. С другой стороны, хранение полных текстов документов в базах данных поисковых систем привело бы, во-первых, к резкому росту их объема, и, во-вторых, поставило бы проблему соблюдения авторских прав.

Для выделения из документа индексационных терминов используются главным образом статистические закономерности распределения частоты появления различных слов в текстах. В частности, в теории индексирования особый интерес представляют явления, поведение которых носит гиперболический характер. Другими словами, произведение фиксированных степеней переменных остается для таких явлений постоянным.

Наиболее известный гиперболический закон, относящийся к статистической обработке текстов, сформулирован Дж. Ципфом. Он касается распределения слов в достаточно больших выборках текста и используется для решения задачи выделения ключевых слов (терминов) произвольного документа.

Рассмотрим некоторый текст, количество слов в котором обозначим как T, а число вхождений каждого слова t_i в этот текст обозначим как n_i. Частота появления слова t_i в таком случае будет определяться формулой:

$$(TF)_i = \frac{n_i}{T}$$

TF — отношение числа вхождений некоторого слова к общему числу слов документа. Если расположить слова текста в порядке убывания частоты их появления, начиная с наиболее часто встречающихся, то произведение частоты слова $(TF)_i$ на порядковый номер частоты будет постоянным для любого данного слова t_i:

$$(TF)_i * r_i = C \qquad (1)$$

где C – некоторая константа, r_i – порядковый номер (ранг) частоты слова.

Выражение (1) описывает функцию вида $y = k/x$ и её график – гипербола, или прямая в логарифмических координатах (рис. 2).

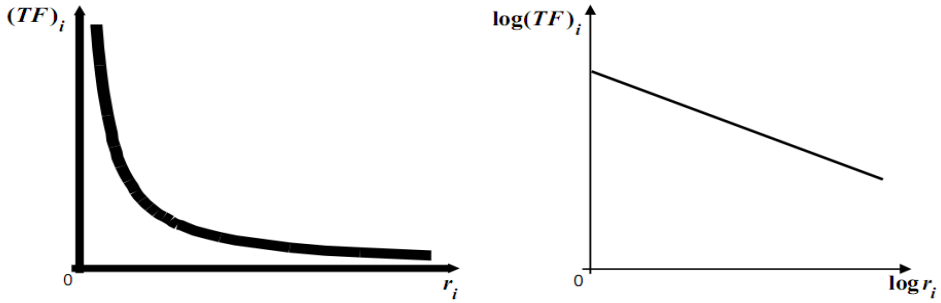

Рис. 2. Закон Ципфа.

Практика показывает, что наиболее значимые слова лежат в средней части графика зависимости (рис. 3). Иными словами, самыми ценными для представления содержания документов являются термины не слишком редкие и не слишком частые. Слова, которые попадаются слишком часто, в основном оказываются предлогами, союзами и т. д. Редко встречающиеся слова также не имеют решающего смыслового значения в большинстве случаев.

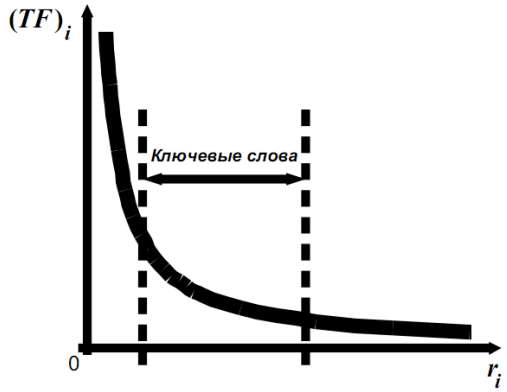

Рис.3. Выделение ключевых слов по закону Ципфа

Границы выделения ключевых слов определяют качество поиска. Высокочастотные термины хоть и не являются специфическими, но все же дают большое число совпадений при сравнении терминов запроса и документа. Тем самым обеспечивается выдача многих релевантных документов, то есть увеличивается полнота поиска. С другой стороны, низкочастотные термины вносят относительно небольшой вклад в поиск нужных документов, так как редкие термины дают малое число совпадений образов запроса и документа. Но если они все же совпадают, то соответствующий найденный документ почти наверняка является релевантным .

Ширина и границы диапазона частот зависят от используемых механизмов поиска, а также от анализируемых документов.

Во всех существующих методах индексирования применяется процедура исключения некоторых высокочастотных терминов, которые заведомо не являются ценными для отражения содержания документа . Для исключения общеупотребительных слов, к которым относятся предлоги, союзы, артикли, вспомогательные глаголы, частицы, местоимения и т. п., используются стоп-словари. Стоп-словарь (стоп-лист, стоп-список, отрицательный словарь) – это словарь служебных и неинформативных терминов, которые не должны входить в число терминов индексации.

8.3. Анализ информационных массивов

Понятие относительной частоты.

Использование во время индексации частоты встречаемости термина в документе (абсолютной частоты) эффективно лишь в случае очень малого объема информационного массива. В действительности же современные массивы данных образованы тысячами и десятками тысяч документов, а в Интернете доступны миллиарды информационных объектов. Поскольку число слов, используемых при индексации, ограничено числом слов в естественном языке и стоп-словарем, для индексации разных документов использовались бы одни и те же термины. Применение абсолютных значений частоты привело бы к резкому уменьшению точности поиска из-за постоянного использования при индексировании высокочастотных терминов, которые встречаются в большинстве документов.

Один из методов усовершенствования этих грубых частотных параметров является метод использования относительных частот терминов в массиве. При этом частота появления термина в данном документе сравнивается с частотой появления этого же термина во всем информационном массиве. Наиболее адекватным при индексации оказывается тот термин, который отражает

содержание отдельного документа и в то же время отличает один документ от другого.

В частотной модели индексирования предпочтительными для описания документов являются те термины, которые встречаются с высокой частотой в отдельных документах, а суммарная частота их появления в массиве низка.

Определим документную частоту термина t_i как число документов массива, в которых встречается этот термин, и обозначим ее $(DF)_i$. Тогда взвешивающую функцию, обратную документной частоте IDF можно определить следующим образом:

$$(IDF)_i = \log \frac{N}{(DF)_i}$$

где N – общее число документов в информационном массиве. Функция приписывает наибольшие значения терминам, появляющимся лишь в нескольких документах. Чем чаще термин встречается в документах массива, тем меньше значение обратной документной частоты.

Несколько иной подход применяется при использовании методов оценки различительной силы термина. Здесь хорошим для индексации считается такой термин, который делает документы максимально непохожими друг на друга. Тем самым обеспечивается максимальное удаление одного документа от другого в пространстве индексирования. Плохим считается такой термин, который делает документы более похожими друг на друга, вследствие чего различить их становится труднее.

Чем больше будет разделение отдельных документов, то есть чем менее похожими будут соответствующие векторы поисковых образов, тем легче будет находить одни документы, отбрасывая другие. Если же документы представлены похожими векторами терминов, пространство индексирования сжимается, и обеспечить достаточное разграничение релевантных и нерелевантных документов затруднительно.

Значимость термина t_i измеряется его различительной силой $(DF)_i$. Она определяется как разность между средним попарным подобием документов, когда термин t_i отсутствует в векторах документов массива, и средним попарным подобием, когда термин t_i присутствует. Если данный термин представляет ценность для индексирования, его присутствие в векторе документа должно делать документы менее похожими друг на друга. Тогда среднее попарное подобие уменьшается, а различительная сила становится положительной. В противном случае значение различительной силы отрицательно.

Распределение частоты встречаемости терминов

Практика показывает, что хорошие, средние и плохие индексационные термины можно характеризовать по распределению их документной частоты $(DF)_i$ и распределению частоты встречаемости F_i. Суммарная частота встречаемости термина t_i в массиве документов определяется следующей формулой:

$$F_i = \sum_{k=1}^{N} (f_i)_k$$

1. Лучшими для индексации терминами с наивысшими значениями различительной силы являются термины со средними значениями суммарной частоты встречаемости F_i и документной частотой, составляющей менее половины его частоты как термина (суммарной частоты в массиве).
2. Следующими по качеству являются термины со значениями различительной силы, близкими к нулю, и очень низкой документной и суммарной частотой.
3. Худшими терминами, имеющими отрицательные значения различительной силы, являются те термины, которые имеют высокую документную частоту (порядка объема всего массива документов) и суммарную частоту термина большую, чем число документов в массиве.

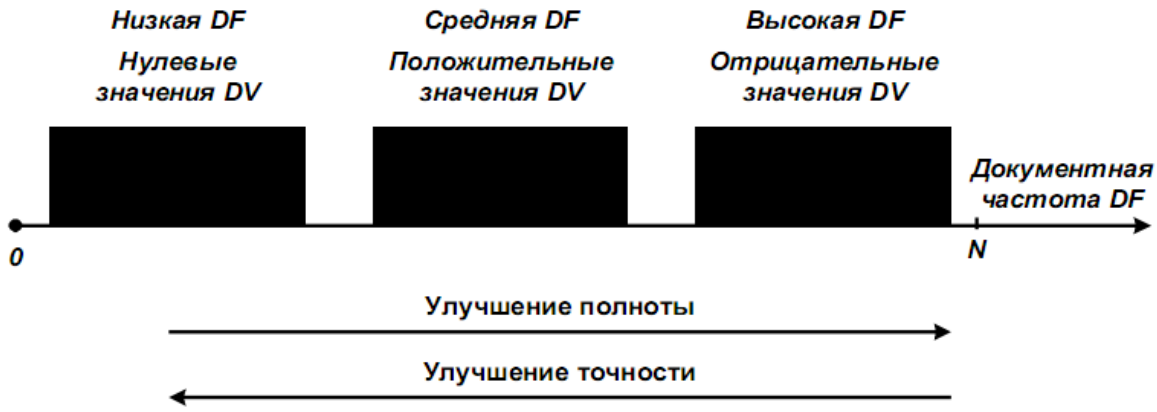

Рис. 4 Характеристика терминов по распределению документной частоты

Рис. 4 иллюстрирует вышеописанное разделение терминов. Если расположить термины в порядке увеличения документной частоты $(DF)_i$, то индексационные термины должны, насколько это возможно, попадать в средний интервал значений.

Внутри каждой из этих категорий, и вообще в массиве документов, термины с относительно плоскими распределениями, для которых частота термина при переходе от документа к документу меняется незначительно, имеют более низкие значения различительной силы. Наоборот, термины с более острыми распределениями, которые часто встречаются в некоторых документах и редко – в остальных, имеют более высокие значения различительной силы. Индексационные термины должны обладать средними по величине значениями документной частоты, и иметь распределения частот, сосредоточенные в одной точке.

На рис. Рис. 5 изображено несколько типичных распределений частот терминов. Наилучшими для индексации являются термины, имеющие распределение (рис. 5а). Они обеспечивают приемлемые значения полноты и точности поиска. Термины с распределениями (рис. 5б) повышают точность, но резко снижают полноту поиска, а с распределениями (рис. 5в) – наоборот, увеличивают полноту, но уменьшают точность. Наконец, равномерное распределение частоты (рис. 5г) свойственно общеупотребительным терминам, которые не обеспечивают ни надлежащей точности поиска, ни его полноты.

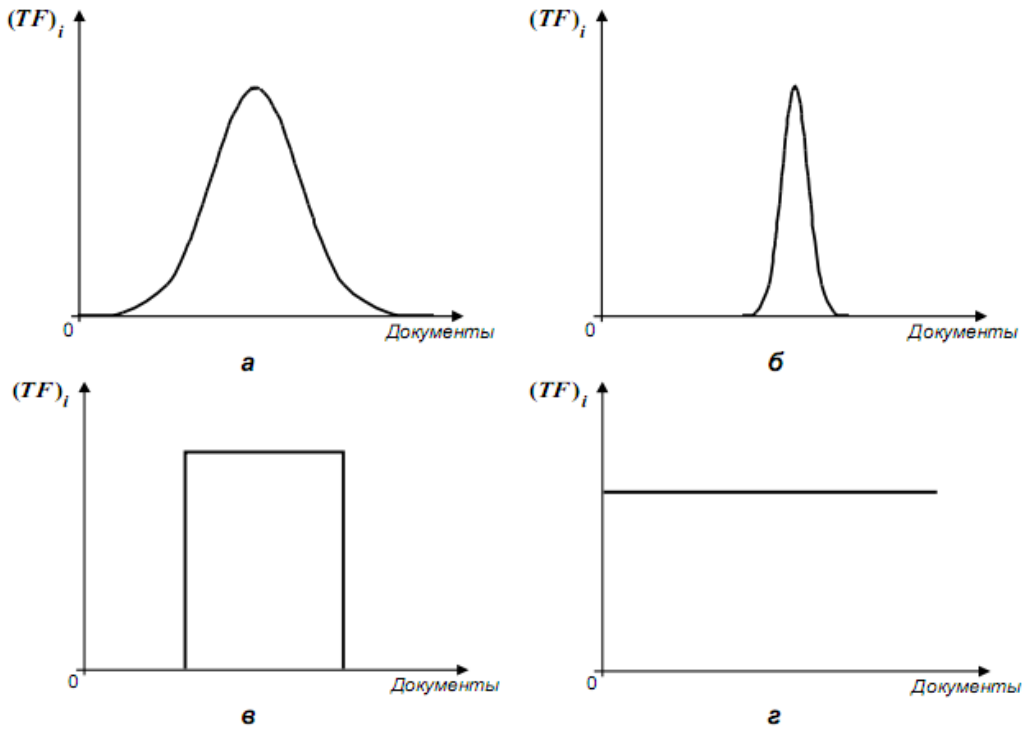

Рис.9. Распределение частот терминов в документе

Определение весов терминов

После того как из документа отобраны термины для поискового образа, возникает вопрос об оценке их значимости для поиска. Ценность того или

иного термина определяется его способностью наиболее адекватно характеризовать содержание документа. Обычно она характеризуется некоторым весовым коэффициентом, который рассчитывается в процессе индексации. Запрос, выражающий информационную потребность пользователя, состоит из отдельных терминов. Во время выполнения поискового алгоритма происходит сравнение терминов запроса и поискового образа документа и определяется степень их близости, то есть формальная релевантность.

Чем больше вес термина в документе, тем более релевантным оказывается этот документ, и тем более высокую позицию этот документ занимает в списке результатов поиска. Особенно актуальным такое упорядочение представляется для крупных информационных массивов.

Таким образом, взвешивание терминов необходимо для решения главной задачи поисковой системы – обеспечения пользователя релевантными документами. Веса также влияют на составление поисковых образов документов. В предыдущих разделах при анализе частот встречаемости терминов был описан ряд критериев, по которым происходит отбор индексационных терминов. Численной характеристикой этих критериев может быть вес терминов. Поскольку количество терминов, которые могут быть использованы для индексации, ограничено, термины, получившие наименьший вес, отбрасываются.

Наиболее простая и самая распространенная модель поиска – булева модель– использует двоичную систему взвешивания терминов. Этот метод реализуется на стадии отбора индексационных терминов, и заключается в том, что терминам, вошедшим в поисковый образ, приписывается единичный вес, а остальным терминам – нулевой вес. Таким образом, все термины из поискового образа документа считаются равнозначными.

Недостатки булевой модели широко известны: использование абсолютных единичных весов приводит к значительным трудностям восприятия результатов поиска, когда в ответ на запрос пользователю система выдает множество неупорядоченных документов, поисковые образы которых содержат термины запроса. Выделение истинно релевантных документов из этого множества представляет значительные трудности.

Выходом из такой ситуации является приписывание терминам дифференцированных весов. Термины поискового образа одного и того же документа в таком случае могут иметь различный вес. Одновременно значение веса для одного и того же термина может быть различным в разных документах.

Частотная модель

Частотная модель взвешивания терминов тесно связана с частотным методом индексирования. Одна из наиболее известных весовых функций записывается следующим образом:

$$\text{TF} - \text{IDF} = W_i = (TF)_i \times (IDF)_i$$

Здесь W_i – вес, приписываемый термину терминati, $(TF)_i$ – частота термина в документе, $(IDF)_i$ – обратная документная частота.

Примеры расчета

Для первого примера расчета $TF - IDF$ возьмём документ из 10000 символов в котором слово «пропан» встречается 25 раз, а коллекция состоит из 2 миллионов документов, в 2000 из которых также встречается данное слово.

$$TF = 25/10000 = 0{,}0025$$
$$IDF = log(2000000/2000) = 3$$
$$TF - IDF = 0{,}0025 \times 3 = 0{,}0075$$

Для второго примера расчета $TF - IDF$ возьмём тот же документ из 10000 символов в котором союз “но” встречается 30 раз, а коллекция по-прежнему состоит из 2 миллионов документов, в 200000 из которых также встречается данное слово.

$$TF = 30/10000 = 0{,}003$$
$$IDF = log(2000000/200000) = 1$$
$$TF - IDF = 0{,}003 \times 1 = 0{,}003$$

Так как мы использовали идентичные документ и коллекцию, можем сравнить семантический вес слов: $0{,}0075 > 0{,}003$, следовательно слово «пропан» имеет больше веса, чем союз «но», что справедливо.

Вероятностная модель

Недостатком частотных методов взвешивания терминов является тот факт, что частотные веса рассчитываются формально, без учета реальных информационных потребностей. Для того чтобы установить соответствие между истинной информационной потребностью и терминами, составляющими поисковый образ документа, разработана вероятностная модель оценки весов терминов.

Вероятностная модель основана на точной оценке вероятности того, что данный документ является релевантным данному запросу .

Обозначим вероятность такого события как $P(w_1|d)$, где w_1 – событие, которое состоит в том, что документ d является релевантным по отношению к запросу q . Для определения вероятности $P(w_1|d)$, воспользуемся теоремой Байеса:

$$P(w_1|d) = \frac{P(d|w_1)P(w_1)}{P(d)}$$

Здесь $P(w_1)$ – вероятность того, что случайно выбранный документ является релевантным, $P(d)$ – вероятность того, что из всего множества документов для рассмотрения выбран документ, $P(d|w_1)$ – вероятность того, что документ d выбран из множества релевантных документов.

Аналогично, предположим, что $P(w_2|d)$ – вероятность того, что документ d окажется нерелевантным.

Латентно-семантический анализ

Основное предназначение взвешивания терминов, как отмечалось выше, заключается в определении того, насколько полно они отражают содержание документа. Как показывает практика, частотные методы оценки весов имеют ряд недостатков. Следствием этого является получение в результате поиска нерелевантных и отсутствие истинно релевантных документов.

Во-первых, рассмотренные методы не учитывают тот факт, что частоты встречаемости различных терминов зависят друг от друга. Термины не появляются в документе независимо от остальных терминов, они могут быть, например, объединены в словосочетания, устоявшиеся обороты и т.п. Другой проблемой является синонимия и полисемия (многозначность).

Под синонимией понимается тот факт, что любое явление или предмет могут быть выражены различными способами. В зависимости от контекста, знаний человека, манеры письма одни и те же сведения описываются разными терминами (синонимами). Например, синонимы «дисплей» и «монитор» определяют один и тот же предмет.

Полисемия, напротив, заключается в том, что большинство слов в языке имеет несколько значений. Один и тот же термин может обозначать абсолютно разные понятия. Соответственно, наличие того или иного термина в некотором документе не означает того, что документ является релевантным запросу, в котором содержится

Например, два документа могут обсуждать одну и ту же тему с использованием разных слов, что может, по-видимому, привести к тому, что программа посчитает, что они никак не связаны друг с другом. Тем не менее, соответствующие релевантные слова обоих документов могут в значительной степени встречаться как в нескольких, так во многих других сходных по тематике документах: эта информация может означать, что все эти слова каким-то образом семантически связаны между собой, поэтому и эти два примера

документов будут потенциально связаны, несмотря на то, что значительно отличаются в словах, которые они содержат.

Описанные проблемы решает латентный семантический анализ (ЛСА), а также известный как латентно-семантическое индексирование (ЛСИ).

Допустим, ставится задача написать алгоритм, который сможет отличать новости о политике от новостей о культуре. Первое, что сразу же приходит на ум, это выбирать слова, которые встречаются исключительно в статьях каждого вида и использовать их для классификации. Очевидная проблема такого подхода: как перечислить все возможные слова и что делать в случае, когда в статье есть слова из нескольких классов. Дополнительную сложность представляют омонимы, т.е. слова, имеющие множество значений. Например, слово «банки» в одном контексте может означать стеклянные сосуды, а в другом контексте это могут быть финансовые институты.

Латентно-семантический анализ производит отображение документов и отдельных слов в так называемое «семантическое пространство», в котором и производятся все дальнейшие сравнения. При этом делаются следующие предположения:

1) Документы — это просто набор слов. Порядок слов в документах игнорируется. Важно только то, сколько раз то или иное слово встречается в документе.

2) Семантическое значение документа определяется набором слов, которые, как правило, идут вместе. Например, в биржевых сводках, часто встречаются слова: «фонд», «акция», «доллар»

3) Каждое слово имеет единственное значение. Это, безусловно, сильное упрощение, но именно оно делает проблему разрешимой.

Латентно-семантический анализ

Латентно-семантический анализ (ЛСА) (англ. Latent semantic analysis, LSA), также известный как латентно-семантическое индексирование (ЛСИ) (англ. Latent semantic indexing, LSI) – это основной метод, используемый для анализа отношений между документами и терминами в коллекции и для извлечения высокоуровневых понятий и преобразования представления документов в соответствии с идентифицированными отношениями.

В общем случае, ЛСА переносит документы коллекции и термины в них в скрытое (латентное) пространство свойств, в котором размерности и измерения идеально соответствуют высокоуровневым понятиям или компонентам. Поэтому, каждый документ представляется в виде взвешенного сочетания таких компонентов, в то время как каждый термин может в различной степени

быть аналогичным образом связан с другими понятиями. Эта схема очень похожа на метод главных компонент (англ. principal component analysis, PCA), который используется для отображения векторного пространства с возможными взаимосвязями между измерениями, на другое пространство, у которого нет таких взаимоотношений.

Пусть дана коллекция из n документов и m различных терминов, извлеченных из них. Для применения модели ЛСА, нужно построить $m \times n$ матрицу «термин-документ» X, с ячейками $x_{i,j}$, содержащими весовые коэффициенты термина t_i в документе d_j . Столбцы матрицы X на практике соответствуют мультимножеству слов (англ. «bag-of-words») для документа, при этом могут быть использованы термины, «взвешенные» по какой-либо схеме: например, это может быть широко-известная схема TF-IDF (от англ. TF – term frequency, IDF – inverse document frequency) или также могут быть эффективны различные схемы, основанные на энтропии.

Внутри данной матрицы для оценки взаимосвязи (корреляции) может быть вычислено скалярное произведение между двумя строками (терминами) или двумя столбцами (документами). Полная матрица взаимосвязей для терминов или документов может быть получена путем вычисления XX^T или X^TX соответственно.

В матрице «термин-документ» применяется сингулярное разложение (англ. singular value decomposition, SVD), математическая методика, которая вычисляет разложение исходной матрицы X на три матрицы.

$$X = U\Sigma V^T$$

Из полученных матриц, матрицы U и V являются ортогональными матрицами размерностей $m \times r$ и $n \times r$ соответственно, а матрица Σ – это $r \times r$ диагональная матрица, содержащая собственные значения. Обоснование заключается в том, что каждое из r собственных значений соответствует одному из вышеупомянутых высокоуровневых компонентов, отслеживаемых в коллекции документов, и обозначает, насколько этот компонент актуален во всей коллекции.

Собственные значения сортируются по диагонали матрицы Σ в порядке убывания, так что те собственные значения, которые идут первыми, связаны с наиболее важными компонентами. Это позволяет легко отрезать наименее важные компоненты до числа $k \leq r$, просто удалив соответствующие строки и столбцы в матрицах. Это сокращение потенциально позволяет удалить «шум»

в данных, который может быть составлен, например, из терминов или групп, появляющихся только в нескольких документах и плохо связанных с другими.

Как только такое значение k установлено, можно рассчитать построенную аппроксимированную версию исходной матрицы «термин-документ» X путем перемножения трех усеченных матриц: результирующая матрица X' будет иметь свой ранг, уменьшенный от r до k. Матрица X' структурно идентична матрице X (ее строки и столбцы являются представлениями для тех же терминов и документов, что и в матрице X), но весовые коэффициенты скорректированы теперь так, что «шум» устранен, и учитаны очевидные взаимосвязи между терминами (или между документами). Например, если два термина ta и tb часто встречаются вместе в документах, то документ содержащий только термин ta из этих двух терминов, будет в любом случае иметь вес для термина tb больше нуля (и наоборот).

Из восстановленной матрицы X' или непосредственно из усеченных матриц, используемых для ее вычисления, сходство между терминами и между документами может быть вычислено в соответствии с скорректированными весовыми коэффициентами, которые в общем случае будут отличаться от соответствующих весов, вычисленных из исходной матрицы. В общем случае, когда должны быть найдены документы, наиболее удовлетворяющие запросу, используется общий подход, в котором запрос представляется, как документ, который должен быть сравнен или сопоставлен с каким-то известным документом. Он должен быть сначала отображен в скрытое (латентное) пространство) для того, чтобы пройти такую же коррекцию значений: эта процедура известна как свертка. В скрытом пространстве можно найти связанные документы, которые не содержат точных слов запроса, но при этом строго соответствуют им.