

Ранжированный реферат поисковой выдачи

Семинар «Поисковые технологии–2010»

Андрей Калинин, Николай Харин
ЗАО «Поисковые технологии»

26 февраля 2010 г.



Методы автоматического реферирования документов

Описание используемого метода реферирования

Примеры

Заключение



Поисковые
технологии

Построение рефератов

1. Вручную выполняется аналитиками, преподавателями, лекторами, журналистами, ...
2. При реферировании учитываются особенности аудитории.
3. Сложность и ценность задачи сопоставима с машинным переводом.
4. Рефераты могут быть индикативными, информативными, критическими.
5. Методы разбатываются с 1958-го года, сейчас активно применяются при генерации поисковой выдачи, адаптируются для устройств с небольшими экранами.



Методы автоматического реферирования документов

Описание используемого метода реферирования

Примеры

Заключение



Основные идеи

- ▶ Выделение важных предложений.
- ▶ Организация предложений в смысловые фрагменты.
- ▶ Выявление семантической структуры текста.
- ▶ Перефразировка (обычно — заполнение шаблона.)
- ▶ Использование машинного обучения для настройки параметров.

Выделение предложений (Luhn, 1958)

1. Выбрать важные для данного документа термины (tf).
2. В каждом предложении пометить отобранные термины и выделить их компактные вхождения:

. . [s . s s . . s] . .

3. Принять вес вхождения длины n с количеством важных слов k за

$$W = \frac{k^2}{n}$$

4. Рассчитать вес предложений по наибольшему весу вхождения важных терминов.
5. Отобрать для реферата нужное количество предложений с максимальным весом.

Дополнительные характеристики предложений (Edmundson, 1969)

1. Учёт месторасположения предложения в тексте или структурной составляющей (L).
2. Словари положительных и негативных слов (C).
3. Важность слов вычисляется с учётом их встречаемости в корпусе документов ($tf \cdot idf, K$).
4. Дополнительные веса для терминов из названия и заголовков (T).

Вес предложения вычислялся как

$$W = \alpha \times L + \beta \times C + \gamma \times K + \sigma \times T$$

Исследования показали, что γ должно быть как можно меньше.

Смысловые фрагменты

- ▶ Реферат из отдельных предложений редко читабелен.
- ▶ Предложения разделяют на зависимые и независимые.
- ▶ Смысловой фрагмент содержит в себе одно независимое предложение и несколько зависимых от него.
- ▶ Использование в реферате смысловых фрагментов увеличивает объём реферата, но и делает его более «гладким».



Машинное обучение (Курьес, 1995)

1. Выбирается набор характеристик предложений (длина, месторасположение, наличие слов из словарей и т.п.)
2. Людям предлагается отобрать важные предложения для реферата.
3. На основании этих данных обучается классификатор.
4. Полученный классификатор применяется для обработки новых документов.

Семантическая структура текста и перефразировка

- ▶ Использование явно заданной структуры документа: количество отобранного материала пропорционально длине структурной составляющей.
- ▶ Рассчитывая на заранее заданную структуры в статье (новость можно разделить на информационный «бэк» и непосредственную новость).
- ▶ Используя средства синтаксического анализа, проследить развитие истории в тексте, выбрать характерные фрагменты текста и составить из них реферат.
- ▶ Перефразировка:
 - ▶ Заполнение заранее подготовленного шаблона.
 - ▶ Удаление уточняющих частей предложения.

Методы автоматического реферирования документов

Описание используемого метода реферирования

Примеры

Заключение



Поисковые
технологии

Общая идея

1. Текст очищается от навигационной обвязки.
2. Выделяются предложения.
3. Оценивается зависимость и информативность предложений.
4. Предложения собираются в смысловые фрагменты.
5. Выделяются значимые термины и словосочетания с весами.
6. Фрагменты кластеризуются по значимым терминам.
7. Оценивается информативность смысловых фрагментов (с учётом запроса).
8. Отбираются фрагменты в реферат: из разных кластеров, в соответствии со структурой документа, наиболее информативные.

Определение независимости предложений

Простой способ:

1. Первое предложение в тексте — независимое.
2. Если в начале предложения есть коннектор (он, это, другой, ...), оно зависимое и связано с предыдущим предложением.
3. Отдельный словарь безусловных коннекторов.
4. Короткие предложения считаются зависимыми.
5. Производится поиск ложных коннекторов (находятся в скобках, описываются шаблонами «кроме того», « — это» и т.п).



Информативность предложений

1. Много слов, начинающихся с большой буквы:

Интернет через мобильный телефон Магазин · Форум ·
Блоги · Доска объявлений Добро пожаловать, Гость.

2. Мало запятых.
3. Слишком короткие слова.
4. Сработал шаблон неинформативного предложения.

Выделение фрагментов

К рассмотрению принимаются:

1. Основные смысловые фрагменты:

- ▶ Первое предложение независимо, остальные зависимы.
- ▶ Все предложения информативны.
- ▶ Укладывается в ограничения по размеру фрагмента и содержит целое число предложений.

2. Дополнительные смысловые фрагменты:

- 2.1 Составляются из оставшихся предложений.
- 2.2 Могут содержать неинформативные предложения.
- 2.3 Содержат просто последовательные предложения.

Смысловые фрагменты могут редактироваться по настраиваемым шаблонам, например:

&Beg В &any стало известно, что



Оценка информативности фрагмента

1. Вес информативных терминов фрагмента.
2. Вес фрагмента относительно запроса.
3. Наличие индикативных терминов.
4. Число слов и количество предложений.
5. Порядковый номер фрагмента в тексте и в структурной составляющей.
6. Размер кластера.



Отбор фрагментов в реферат

Объём реферата регулируется:

1. Длиной в символах.
2. Коэффициентом сжатия.
3. Количеством смысловых фрагментов.

Приоритет фрагментов:

1. Фрагменты, имеющие ненулевой вес относительно запроса:
 - 1.1 Основные фрагменты с целым числом предложений.
 - 1.2 Основные фрагменты с неполным последним предложением.
 - 1.3 Дополнительные фрагменты, включающие неинформативные предложения.
 - 1.4 Прочие дополнительные фрагменты.
2. Фрагменты, не содержащие в себе запроса в том же порядке.

Выбор фрагментов: равномерный, весовой, комбинированный.



Ранжированный реферат поисковой выдачи

1. Отбираются документы для составления реферата (с учётом групп).
2. Из документов выделяются смысловые фрагменты, содержащие запрос.
3. Вычисляются веса фрагментов, они кластеризуются, из каждого кластера выбирается один фрагмент с максимальным весом.
4. Ранжированный реферат составляется из заданного числа фрагментов, взятых по убыванию их веса.

Настройка

1. 80 настроечных параметров (не считая параметров алгоритма расчёта весов терминов).
2. 5 двуязычных словарей:
 - ▶ Коннекторы (3000 входов).
 - ▶ Безусловные коннекторы.
 - ▶ Лже-коннекторы (500 входов).
 - ▶ Индикативные выражения (1000 входов, может подбираться под тематику).
 - ▶ Стоп-словари.
3. Шаблоны редактирования предложений (1500 входов).

Методы автоматического реферирования документов

Описание используемого метода реферирования

Примеры

Заключение



Поисковые
технологии

Реферат документа, найденного по запросу

- ▶ Поиск для молодых родителей, запрос «атопический дерматит».
- ▶ Найдена статья на сайте U-Мата, размер после удаления навигационной обвязки 1750 слов (2200 с пунктуацией).
- ▶ Выделено 133 информативных предложения, 69 из которых независимы.
- ▶ Получен 31 смысловой фрагмент.
- ▶ Фрагменты объединились в 19 кластеров, в самом длинном кластере 6 фрагментов, 13 кластеров состоит из одного фрагмента.
- ▶ Отобрано 3 фрагмента.

Атопический дерматит: как помочь малышу? (часть 1) / Детское здоровье

Публикации » Все о детях » Детское здоровье » Авторские статьи

Атопический дерматит: как помочь малышу? (часть 1)

Как современные дерматологи диагностируют и лечат атопический дерматит у детей? Реально ли предупредить развитие заболевания? Чем может помочь традиционная медицина и другие альтернативные методики?

Статья написана на основе материалов, любезно предоставленных главным детским дерматологом г. Екатеринбурга И.Г. Лаврик, а также с учетом информации с форумов и личного опыта автора по лечению данного заболевания у своего старшего ребенка.

Атопический дерматит сегодня считается одним из самых распространенных аллергических заболеваний. По данным исследований, проведенных в России и за рубежом, атопическим дерматитом страдают 30-40% детей. Он начинается, как правило, в первый год жизни и имеет склонность к хронизации или течению с

Выделенные термины

аллергических заболеваний
атопического дерматита ребенка
атопический дерматит
ребенка
профилактика atopического дерматита
профилактика дерматита
дерматита аллергическое
заболевания
вызывают аллергическую реакцию
малышу
петрушки
мамы
реакцию
детях
овощи
сельдерея
кожей

аллергическую реакцию
дерматит
дерматита
аллергических
аллергенов
аллергия
ребенка
кожей ребенка
лечению
кожных
пыльца
капусты
атопический
родители
зуд
организм
укроп

Атопический дерматит

Атопический дерматит сегодня считается одним из самых распространенных аллергических заболеваний. По данным исследований, проведенных в России и за рубежом, **атопическим дерматитом** страдают 30-40% детей. Он начинается, как правило, в первый год жизни и имеет склонность к хронизации или течению с частыми обострениями.

Атопический дерматит – заболевание, приносящее множество неприятностей не только больному ребенку, но и всей его семье. В первую очередь, это зуд кожных покровов, делающий ребенка нервным и нарушающий нормальный сон, и кожные высыпания на открытых участках тела (лицо, руки).

В последние годы у детей чаще стала встречаться аллергия на бананы, киви, авокадо, хурму, гранаты. Употребление сельдерея, петрушки, лука в сыром виде, квашеной капусты, пряностей, мясных и куриных бульонов усиливает проявления **атопического дерматита**.

Ранжированный реферат поисковой выдачи

- ▶ Поиск для молодых родителей, запрос «атопический дерматит».
- ▶ Построена поисковая выдача, сгруппированная по принадлежности к одному сайту, выделено по три документа из группы, взято 10 групп, тем самым реферат построен по 30 документам.
- ▶ Общий объём 32 тысячи слов, 400КБ.
- ▶ 2501 информативное предложение, из них 1708 независимых.
- ▶ Выделено 113 фрагментов, сгруппированных в 42 кластера.
- ▶ В реферат отобрано 4 смысловых фрагмента.
- ▶ Будет приведено два примера: по всем сайтам поиска и с исключением форумов.



Атопический дерматит, чтоб его! (том 5) (форум sibmama.ru, автор Мелания)
Я не против нетрадиционной медицины, я пытаюсь объяснить, что в нашем конкретном случае атопический дерматит сопутствующее заболевание, как следствие Целиакии(т.к. вследствие снижения усваиваемости продуктов, нарушения обменных процессов и т.д. организм склонен к аллергическим проявлениям в том или ином виде).

Атопический дерматит: как помочь малышу? (часть 2) / Детское здоровье (u-mama.ru)

Личный опыт автора и опыт многих родителей подтверждает высокую эффективность лечения атопического дерматита с помощью гомеопатии. Главное здесь – найти «своего» врача, потому что в гомеопатии, как, пожалуй, нигде больше, важно тесное сотрудничество и кропотливая совместная работа пациента и доктора.

От атопического дерматита до бронхиальной астмы у детей - Лечащий врач 01/2006 (медицинский журнал)

Одним из факторов высокого риска развития БА у детей считается атопический дерматит, который можно считать первым (по срокам возникновения) аллергическим заболеванием, а также начальным этапом «атопического марша»: атопический дерматит — аллергический ринит — БА или атопический дерматит — БА — аллергический ринит/БА [3].

Атопический дерматит: как помочь малышу? (часть 1) / Детское здоровье (u-mama.ru)

Атопический дерматит сегодня считается одним из самых распространенных аллергических заболеваний. По данным исследований, проведенных в России и за рубежом, атопическим дерматитом страдают 30-40% детей. Он начинается, как правило, в первый год жизни и имеет склонность к хронизации или течению

Атопический дерматит: как помочь малышу? (часть 1)

Атопический дерматит сегодня считается одним из самых распространенных аллергических заболеваний. По данным исследований, проведенных в России и за рубежом, атопическим дерматитом страдают 30-40% детей. Он начинается, как правило, в первый год жизни и имеет склонность к хронизации или течению с частыми обострениями.

От атопического дерматита до бронхиальной астмы у детей - Лечащий врач 01/2006

Одним из факторов высокого риска развития БА у детей считается атопический дерматит, который можно считать первым (по срокам возникновения) аллергическим заболеванием, а также начальным этапом «атопического марша»: атопический дерматит — аллергический ринит — БА или атопический дерматит — БА — аллергический ринит/БА [3].

FAQ по уходу за кожей новорожденных.

Атопический дерматит - это аллергическое заболевание кожи – аллергодерматоз. Современное определение атопического дерматита звучит так — это хроническое заболевание кожи, в основе которого лежит аллергическое воспаление и ее гиперреактивность, нарушающие естественную реакцию кожи на внешние и внутренние раздражители.

Атопический дерматит: почему?

В настоящее время этиология атопического дерматита (АД) неизвестна, но выделяют 3 группы причинных факторов, вызывающих развитие АД. Таковыми являются аллергены, раздражители (псевдоаллергены) и возбудители инфекций (бактерии, вирусы, грибы).

Методы автоматического реферирования документов

Описание используемого метода реферирования

Примеры

Заключение



Поисковые
технологии

Выводы

- ▶ Реферирование одного документа не всегда получается хорошо (поисковая система возвращает сложные для реферирования документы).
- ▶ Ранжированный реферат поисковой выдачи выглядит значительно лучше, т.к. отбирает фрагменты из большего количества документов.
- ▶ Ранжированный реферат является надстройкой над поиском, выполняющим свою оценку релевантности.



Выводы

- ▶ Реферирование одного документа не всегда получается хорошо (поисковая система возвращает сложные для реферирования документы).
- ▶ Ранжированный реферат поисковой выдачи выглядит значительно лучше, т.к. отбирает фрагменты из большого количества документов.
- ▶ Ранжированный реферат является надстройкой над поиском, выполняющим свою оценку релевантности.

Развитие:

- ▶ Подбор параметров средствами машинного обучения.
- ▶ Академически точная оценка качества рефератов.
- ▶ Ускорение работы системы ранжированного реферирования.