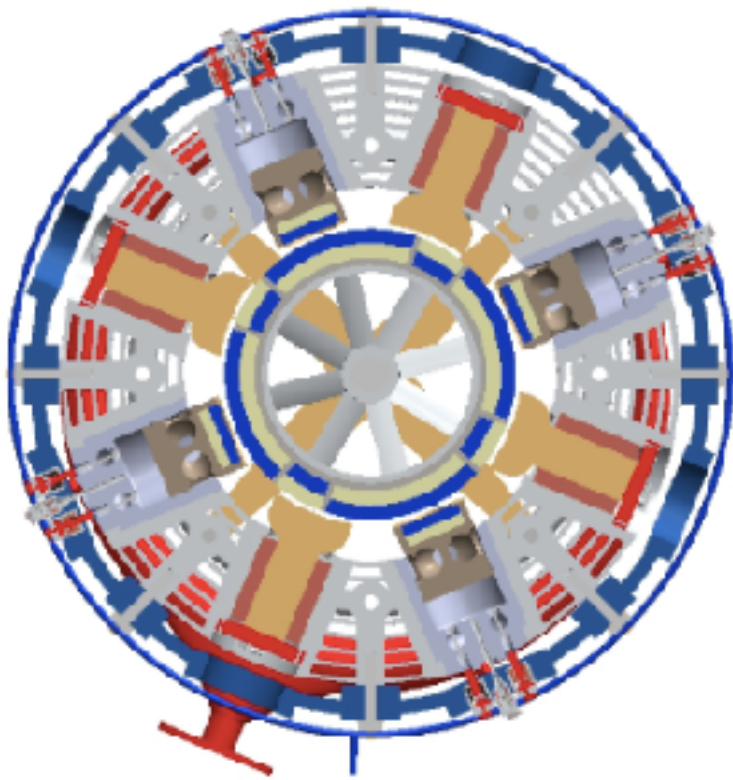


Что значит "Контент-король"?

Кокшаров Сергей

Devaka.ru

Почему "Контент — король"?



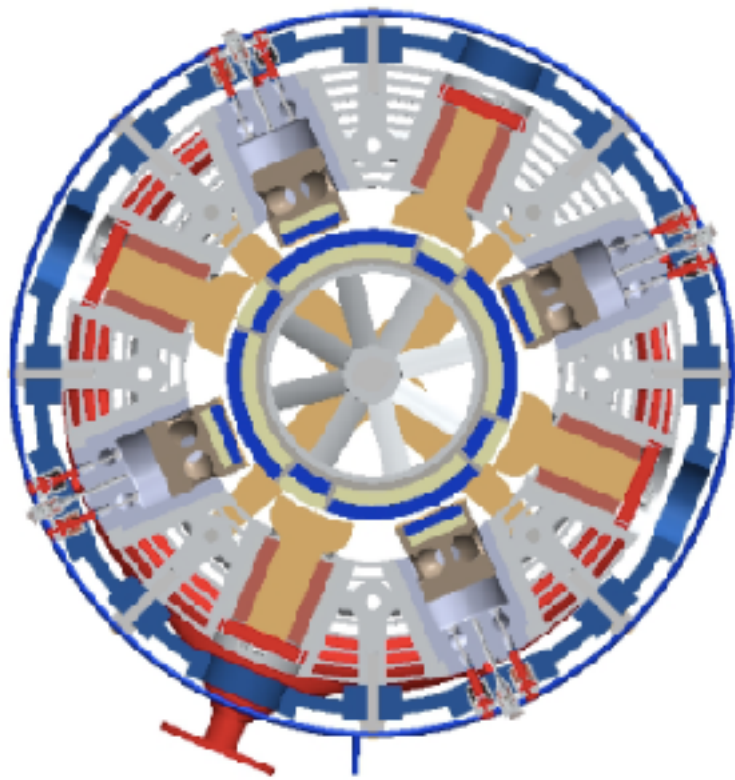
**Текстовые
характеристики**

**Пользовательский
опыт**



Доверие

Разберем немного патентов



**Текстовые
характеристики**

Качество

Новизна

Взаимосвязь слов

Богатство семантики

Авторство

....

Ситуация 1

Создаем сайт, наполняем контентом.
Какие у него должны быть характеристики?
Что такое качественный контент?

Качество контента

Анализ контента для определения спамных страниц

<https://patents.google.com/patent/US20060184500A1/en>

Формирование сигнала качества сайта

<https://www.google.com/patents/US8442984>

Определение мусорного контента на сайтах

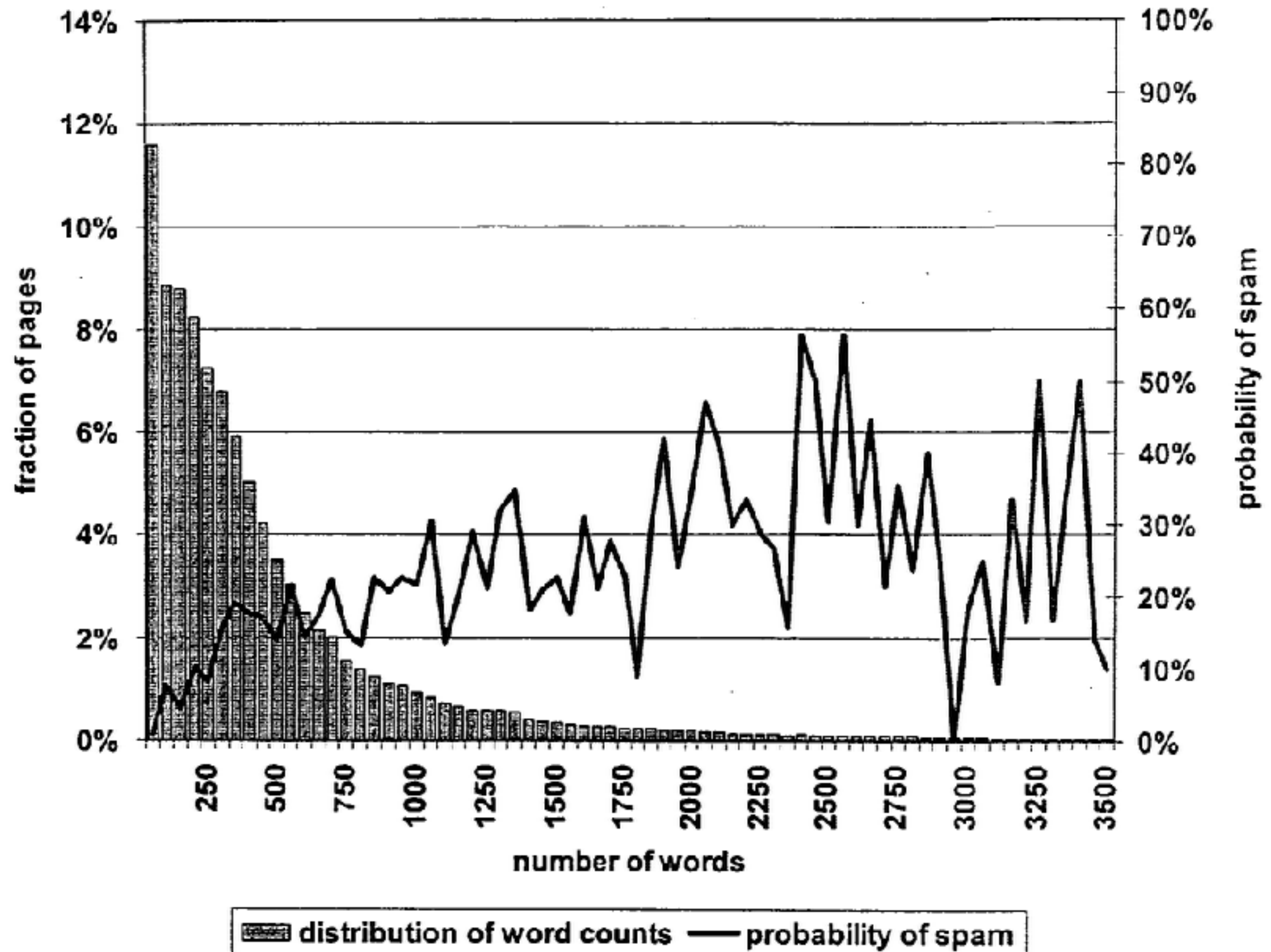
<https://patents.google.com/patent/US8554769B1/en>

- Классификатор вебспама на основе статанализа параметров контента
- Анализируется не только текст страниц, но и заголовки и тексты ссылок
- Второй подход: n-граммный анализ. Кол-во n-грамм, процент страниц с такой n-граммой на сайте.
- Анализируется фразовая модель на основе n-gram и сравнивается с уже известными сайтами, на основе чего формируется quality score

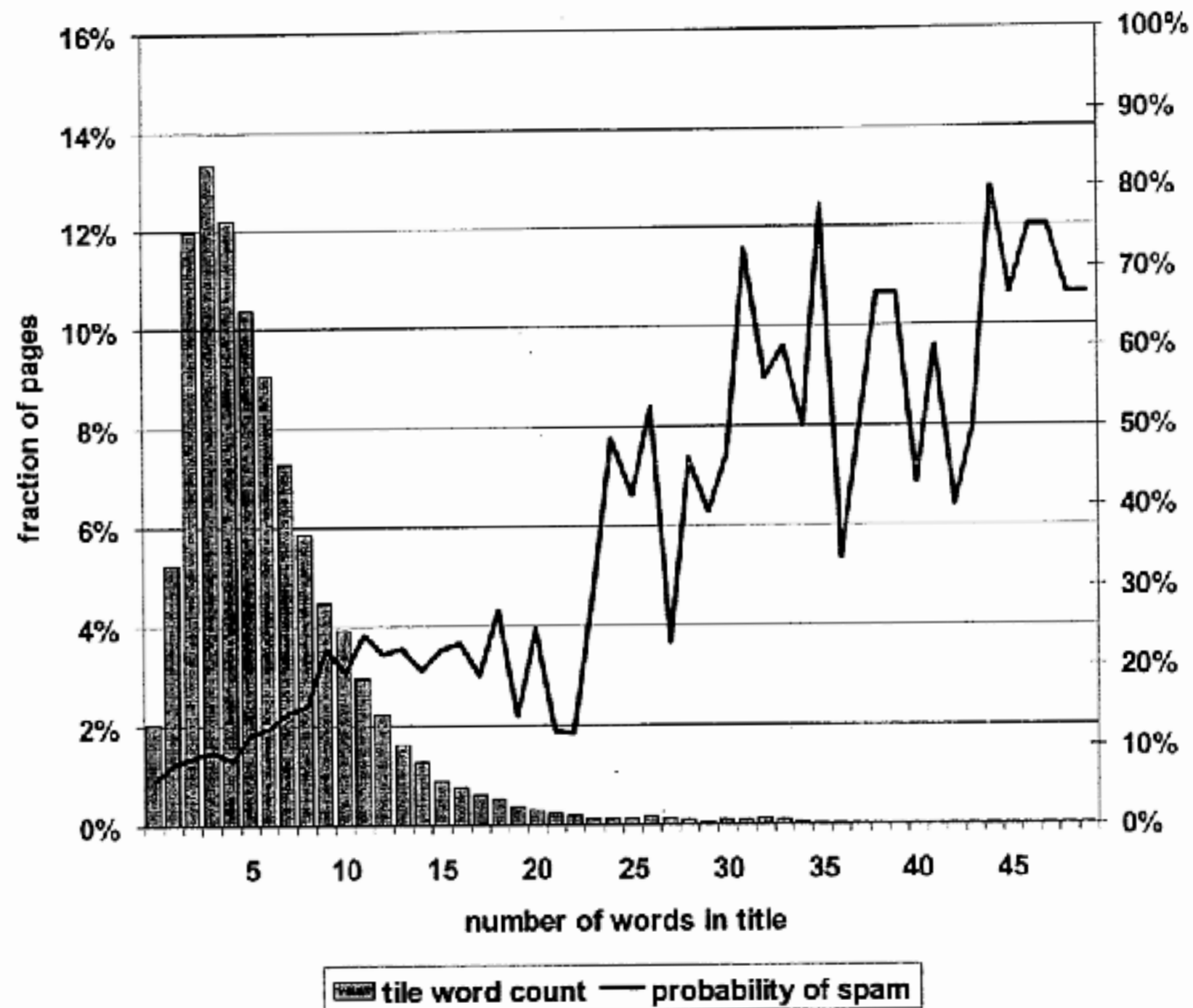
Страница может быть распознана как спам, если:

- имеет большое количество длинных или коротких слов
- большое количество слов
- низкое количество общих слов
- отношение html-кода к видимому тексту
- повторяющийся контент
- большое количество популярных n-gram
- отношение анкорных слов ко всем словам на странице
- и др.

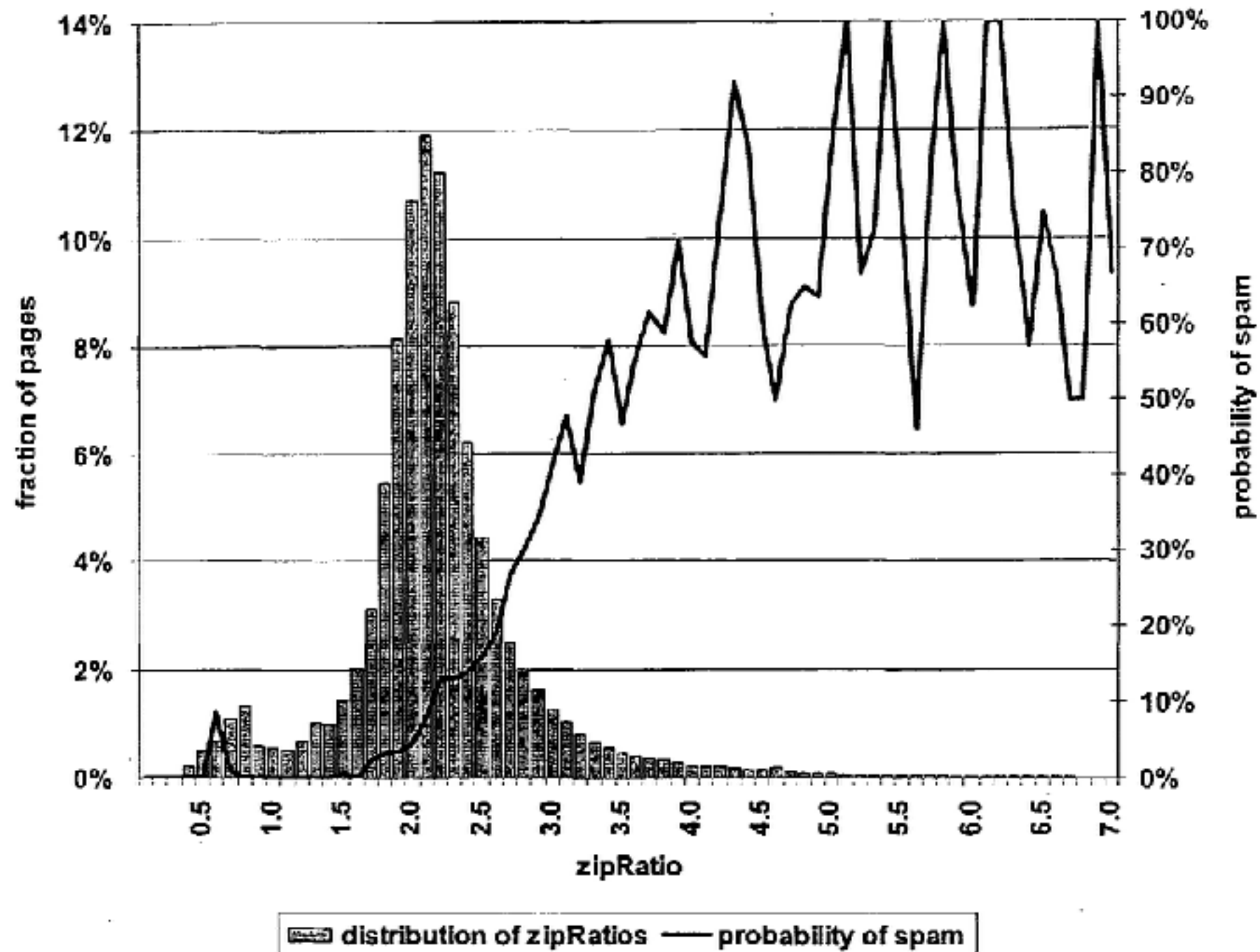
Distribution of Word Counts



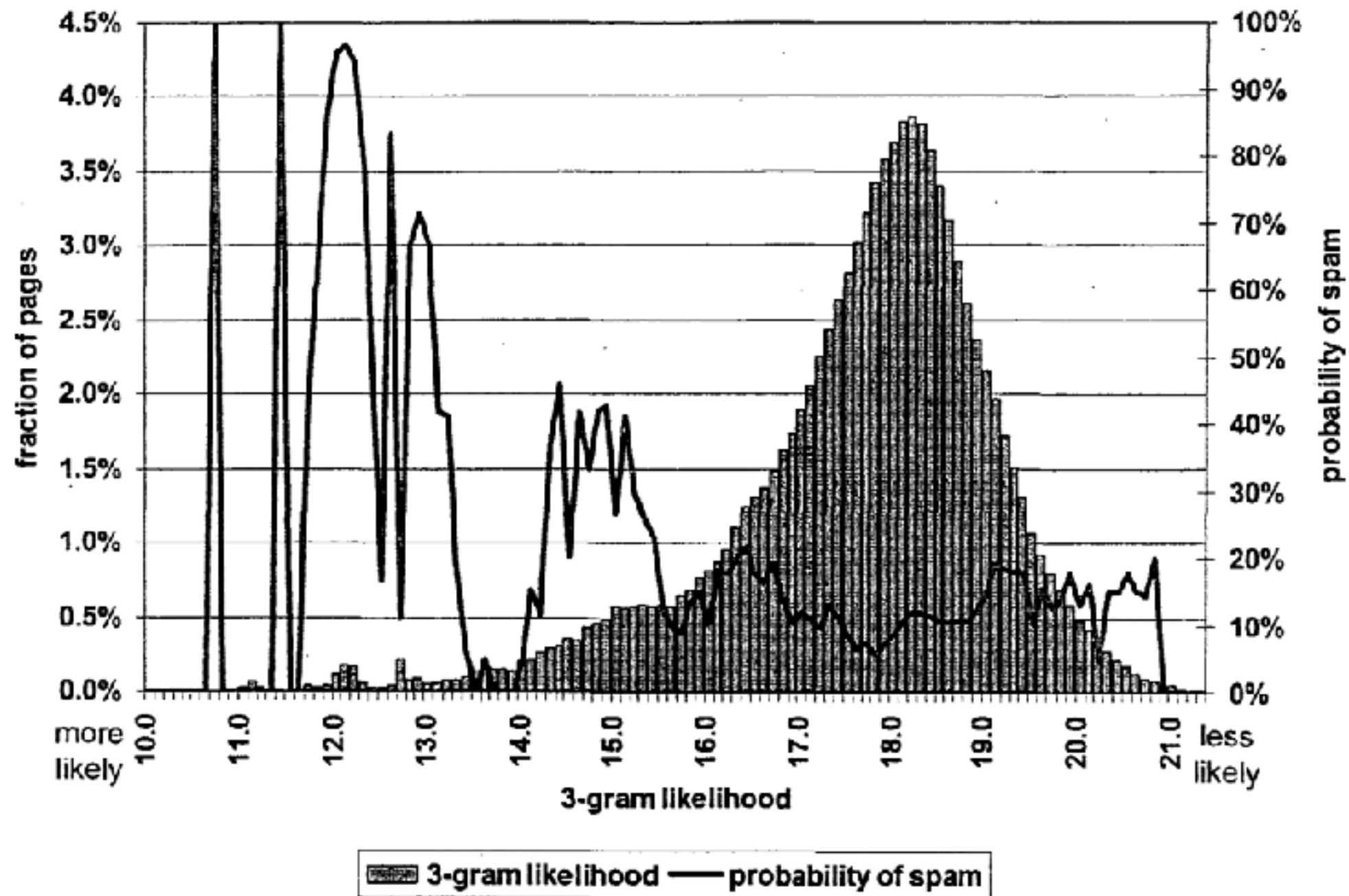
Distribution of Word Counts in Title



Distribution of zipRatios



Distribution of 3-gram Likelihoods (Independent)



Ситуация 2

Редактируем сайт. Хотим сменить дизайн и верстку. Можно ли обновлять контент?

Обновление контента

Поиск информации с учетом исторических данных

<https://www.google.com/patents/US7346839>

- Гугл хранит историю: трафик, поведение, история домена, рейтинг, уникальные слова и биграмы, фразы в анкорах, кол-во новых пассажиров и пр.
- Документы с разной частотой обновления могут ранжироваться по-разному
- Незначительные измерения: яваскрипт, баннеры, навигация, теги.
- Важные изменения: тайтл, анкорный текст исх. ссылок
- Сравнивается средняя дата обновлений для страниц со средней датой обновлений документов по теме.

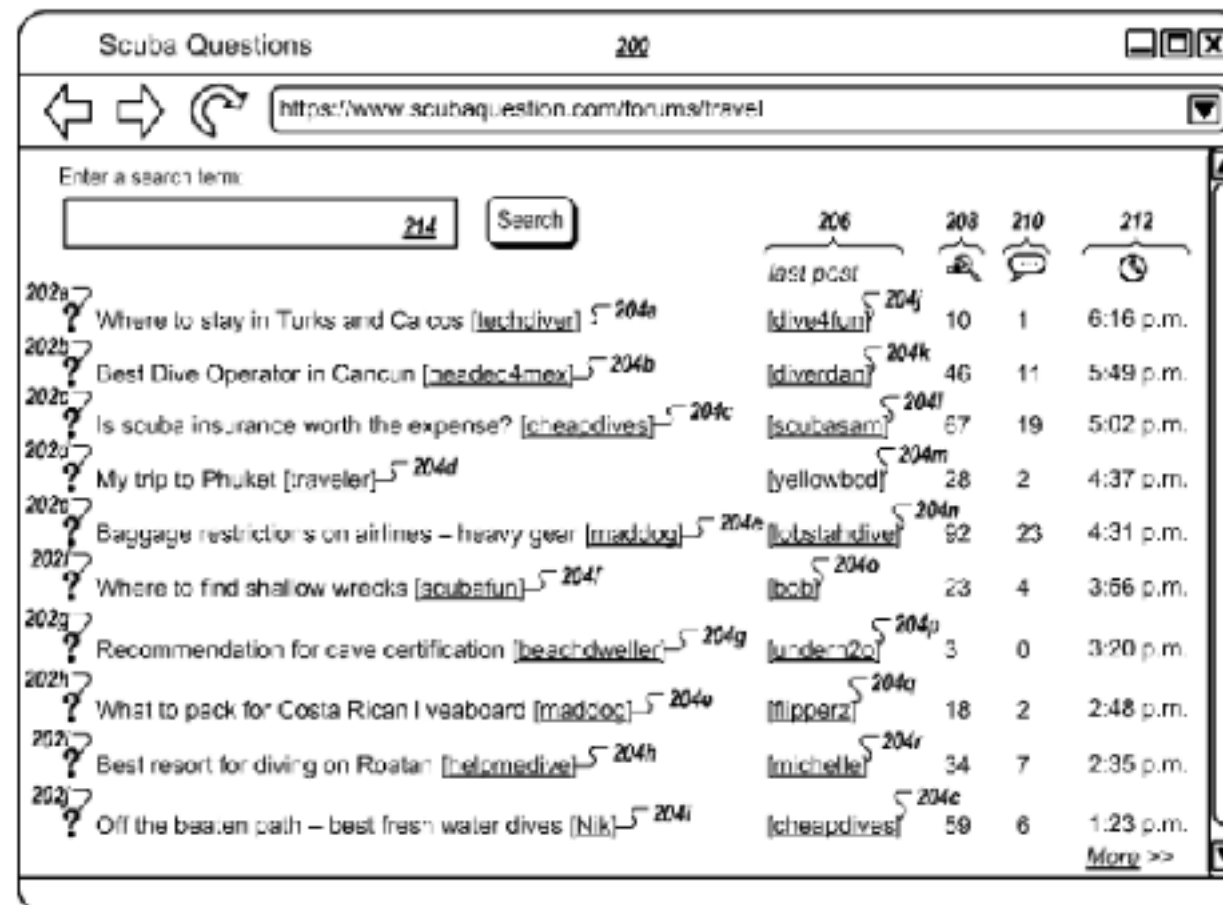
Ситуация 3

На сайте много UGC-контента.
Он ранжируется также, как основной?
Главное побольше текста и ключей?

UGC-контент

Ранжирование UGC-контента

<https://patents.google.com/patent/WO2011050495A1/>



- UGC-сайты: форумы, сайты вопросов-ответов, блоги, соцсети, фото и видео-сайты, где пользователи постят.
- Контент ранжируется в зависимости от рейтинга пользователя, а не только от релевантности
- user credential score - взаимодействие пользователей и их вклад в проект (своего рода ТИЦ)
- contributiveness scores - по графу активности и рейтинга каждого пользователя
- authority scores - качество ответов, можно находить спам

- Учитывается персонализация. Если с кем-то уже взаимодействовали, то рейтинг его объектов повышается
- Качественный ответ на хороший вопрос позитивно влияет на рейтинг
- Плохой ответ влияет отрицательно
- Ответ на вопрос от пользователя с плохим рейтингом также может отрицательно повлиять на рейтинг
- На минутку: у гугла есть гугл+ и ютуб

Ситуация 4

На сайте много таблиц и списков/меню.
В каждом пункте меню нужно ставить ключевое слово, или достаточно в нескольких?

Семантическое расстояние

Ранжирование документов на основе семантического расстояния между терминами документа

<http://www.google.tl/patents/US8606778>

Saturn Facts

- One Orbit of Sun is 10,759.2 Days
- Rotation Period is 10 Hours, 13 Min
- Mass is 95 times that of Earth
- Volume is 744 times that of Earth
- Distance from Sun is 1,429 million kilometers

- Используется не визуальная близость между терминами, а семантическая, зависит от структуры документа
- Определяются списки, независимо от того, были они размечены или нет. То же самое с заголовками
- Заголовки это наиболее близкие термины к тем, что расположены ниже по дереву
- Близость терминов учитывается при расчете релевантности

```
<div align = center>  
<font size=16> Saturn Facts </font>  
  <br>  
  <b> One </b> Orbit of Sun  
  is 10,759.2 Days <br>  
  <b> Rotation </b> Period is  
  10 Hours, 13 Min <br>  
  <b> Mass </b> is 95 times  
  that of Earth <br>  
  <b> Volume </b> is 744 times that  
  of Earth <br>  
  <b> Distance </b> from Sun is  
  1,429 million kilometers <br>  
</div>
```

Ситуация 5

Что по поводу синонимов и LSI-фраз?
Это вообще работает?

LSI-контент

Ранжирование документов, основанное на взаимосвязи слов

<https://www.google.ch/patents/US7996379>

Оценка заменяемых слов

<https://www.google.com/patents/US8504562>

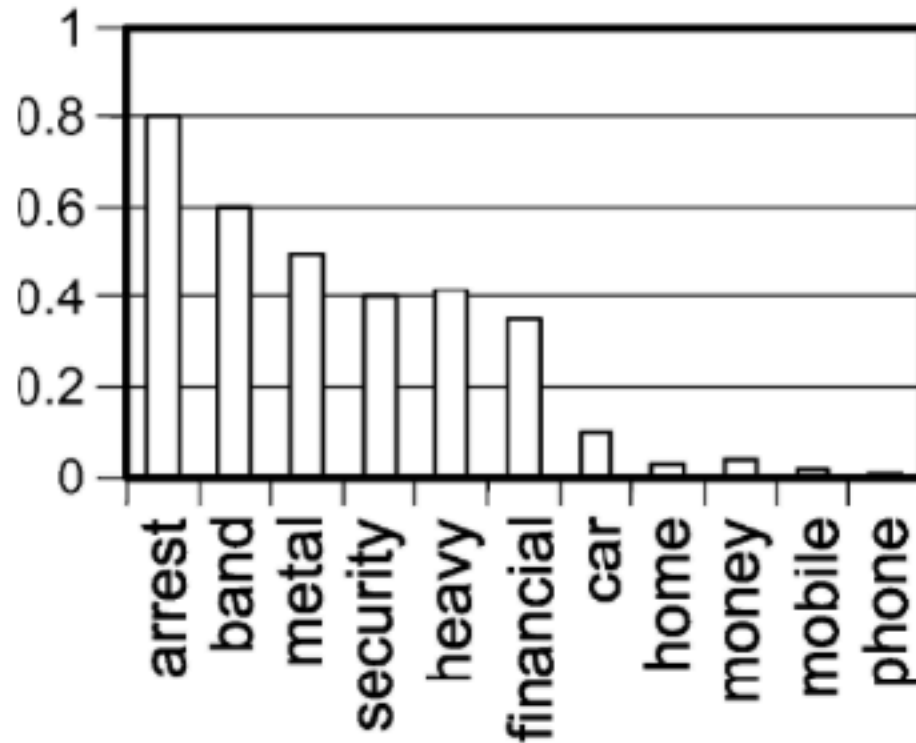
- Google находит синонимы в рамках контекста и расширяет ими запросы
- Для поиска синонимов анализируется топ документов
- Часто встречающиеся термины могут рассматриваться как связанные
- Формируется база знаний
- Запрос расширяется на substitute terms

Queries with "warrant"

arrest warrant
warrant band
warrant heavy metal
warrant financial security

202

warrant



210

a

0.8 arrest
0.6 band
0.5 metal
0.4 security
0.42 heavy
0.35 financial
0.1 car
0.03 home
0.04 money
0.02 mobile
0.01 phone

212

b

0.1
0.05
0.02
0.2
0.05
0.6
0.9
0.6
0.5
0.35
0.3

214

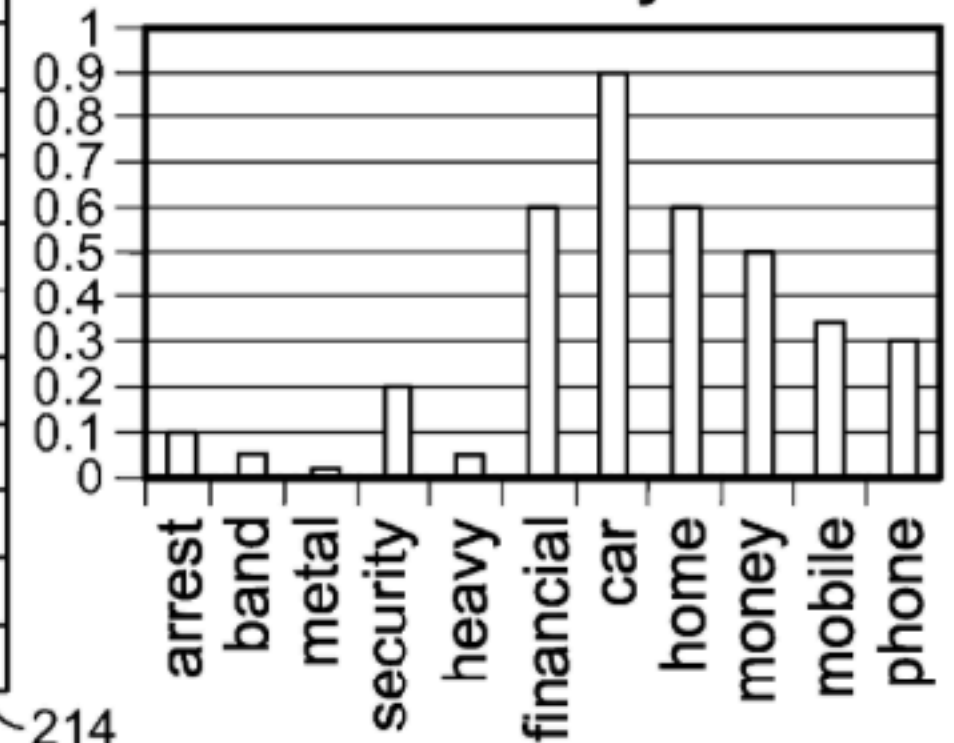
$$\text{Sim}(a, b) = 0.303$$

Queries with "warranty"

used car warranty
home warranty
warranty money back
warranty for mobile phone

204

warranty

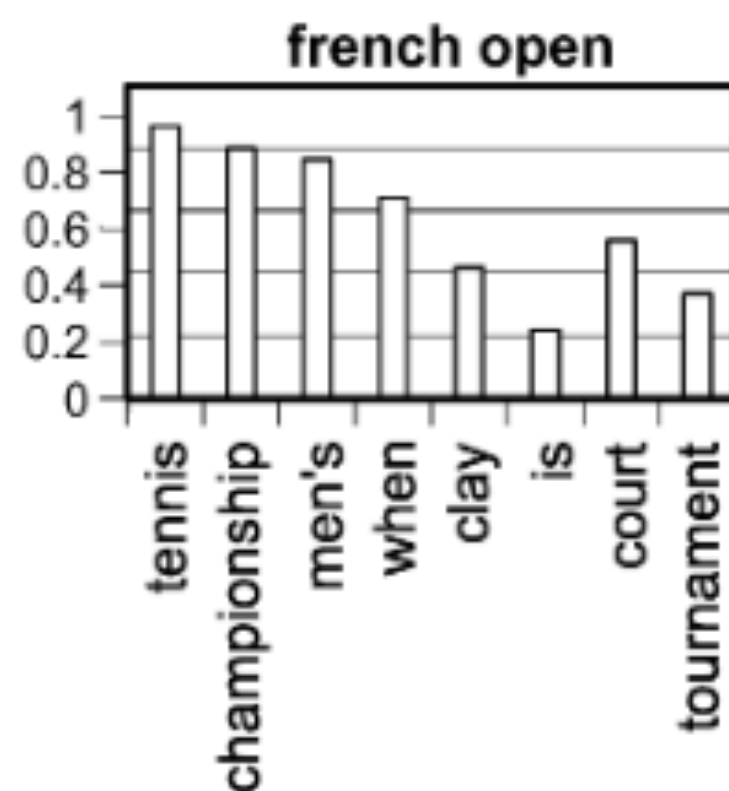


220

∴ "warrant" is not a good substitute term for "warranty"

Queries with "french open"
french open tournament
when is the french open
men's french open
french open championship
french open clay court

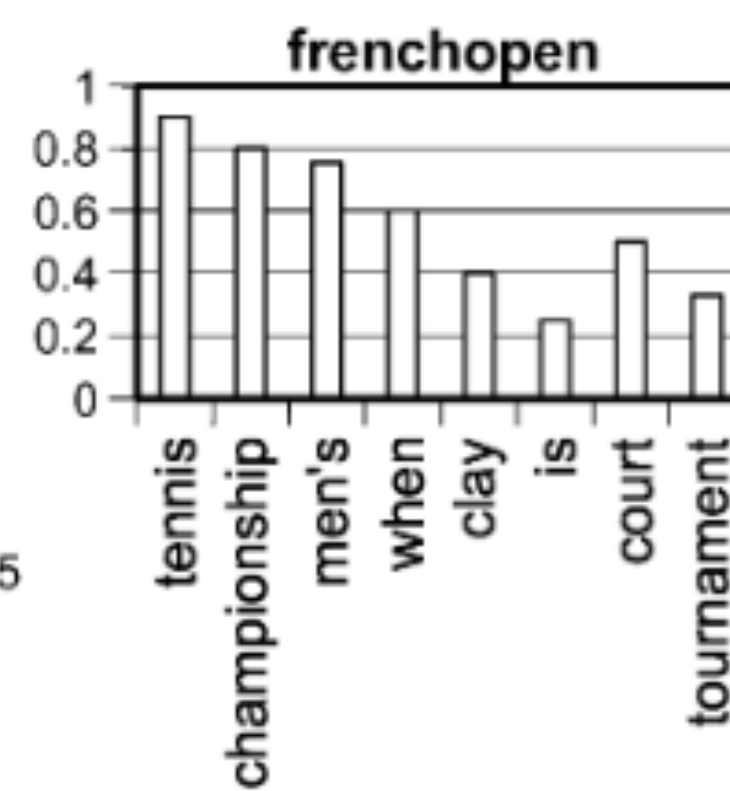
206



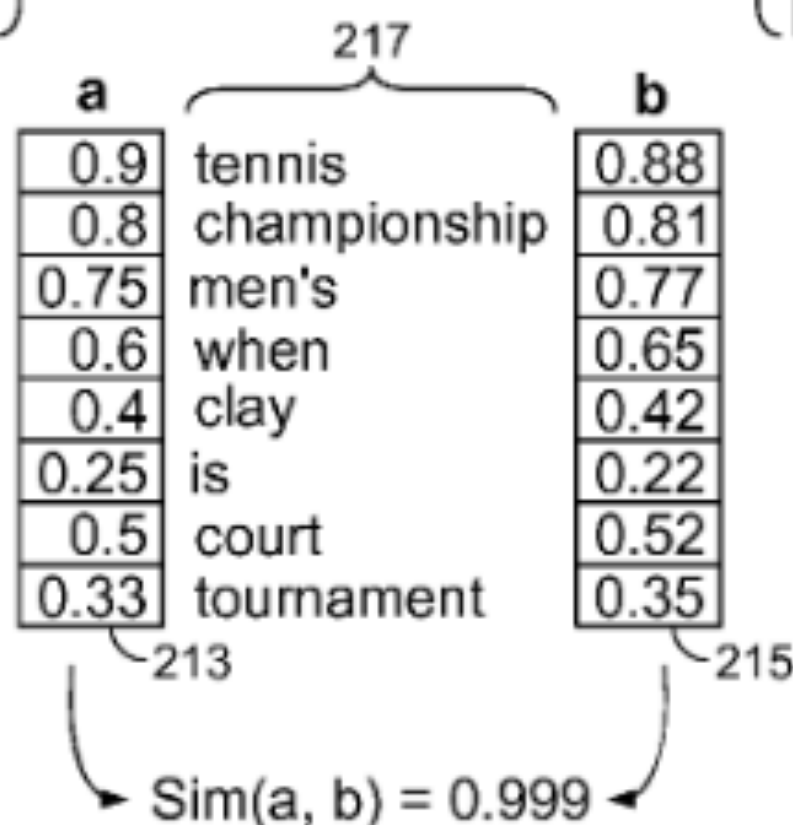
230

Queries with "frenchopen"
frenchopen tennis
frenchopen championship
men's frenchopen
frenchopen clay court

208



240

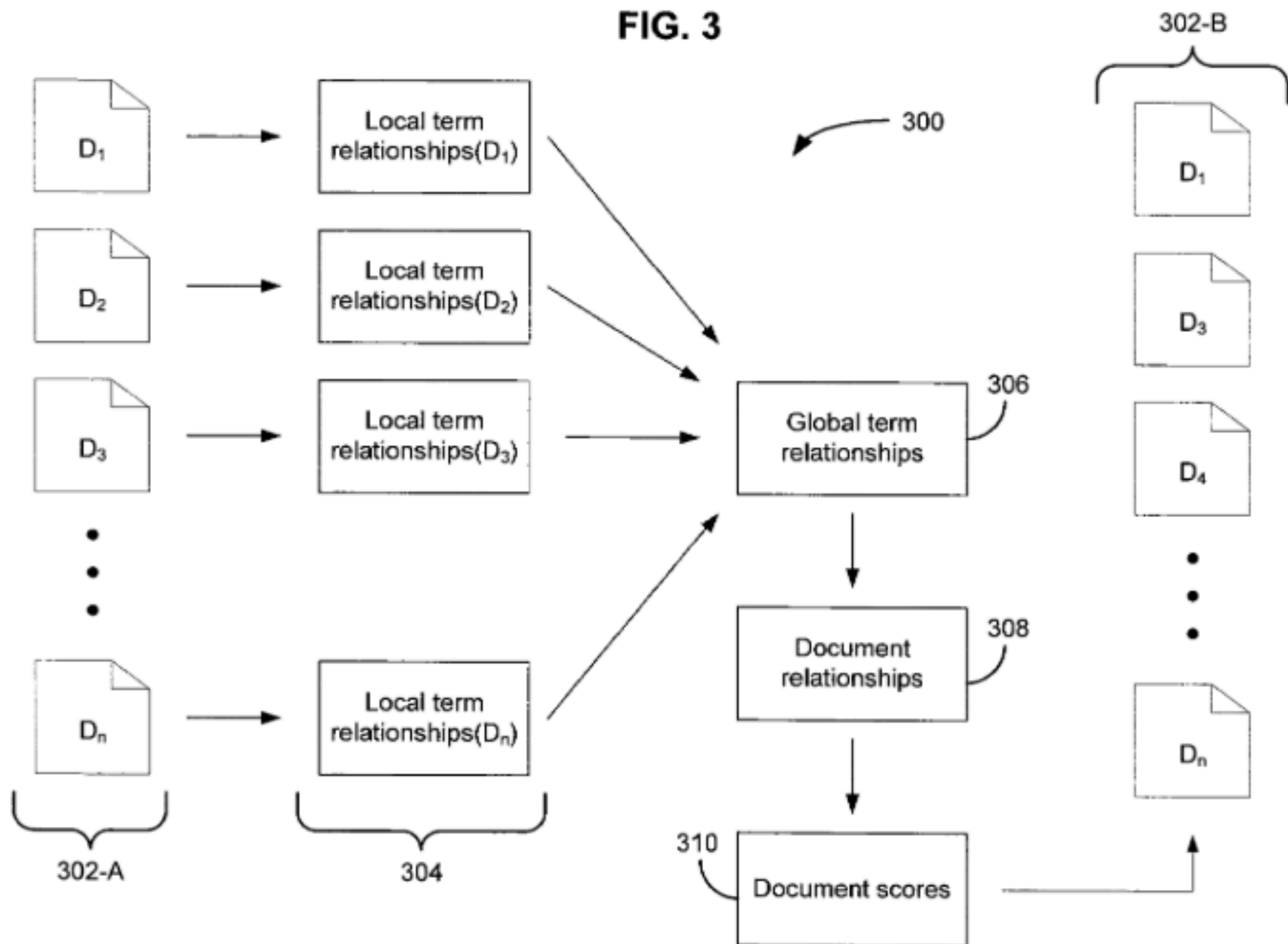


- "french open" is a good substitute term for "frenchopen"

FIG. 2B

- Рассчитываются локальные связи между терминами, находятся наиболее важные термины и их веса.
- Существительные и объекты более весомы, чем глаголы или другие части речи.
- Термины фильтруются. Весомость верхних слов выше.
- Рассчитываются глобальные связи с другими документами, строится граф связей.
- На основе графа рассчитываются веса (по модели PR).

FIG. 3



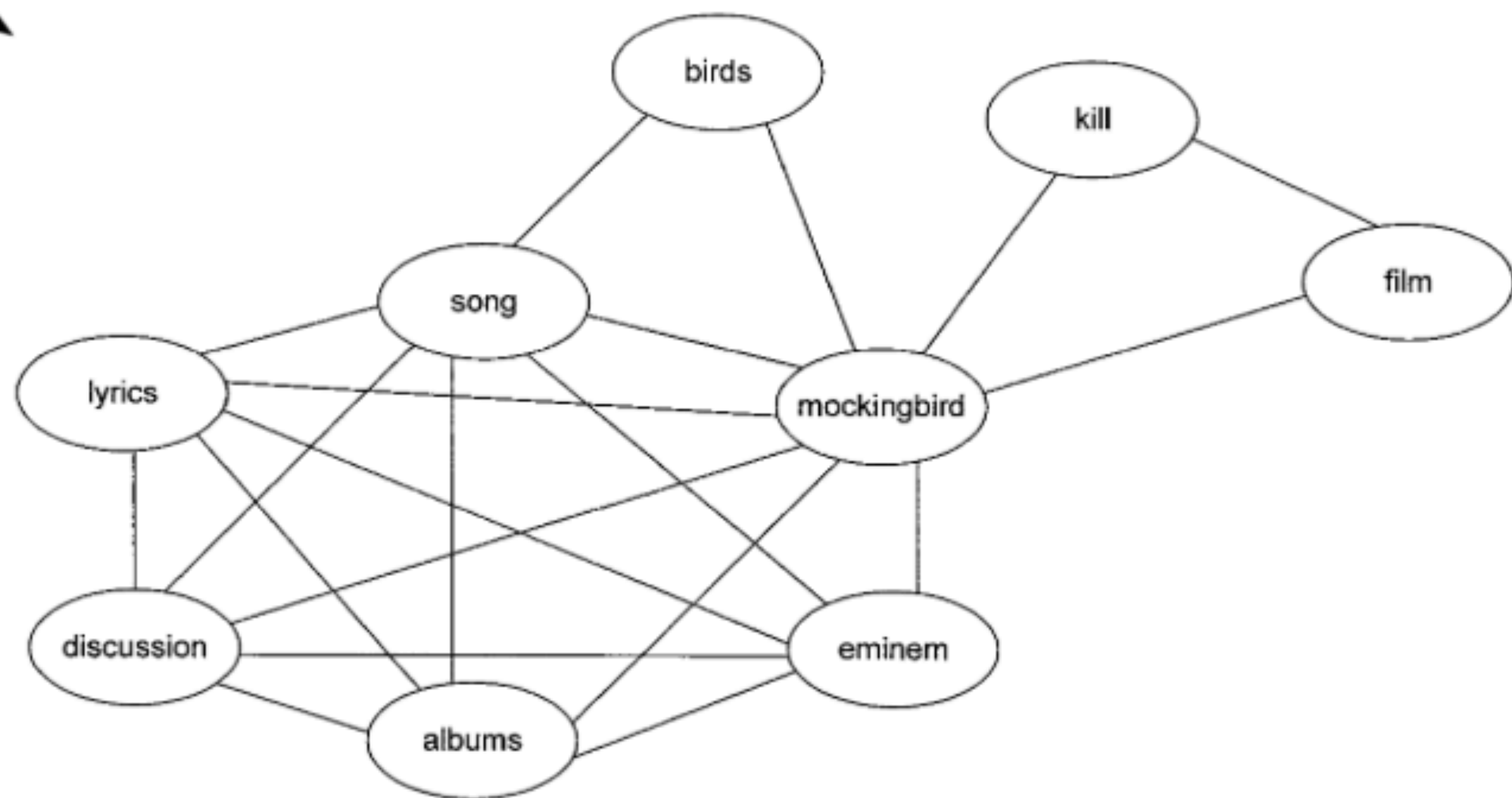


FIG. 5B

Заключение

- Используйте информацию из патентов при работе с сайтами
- Изучайте новые патенты, там много фишек
- Экспериментируйте! Так как какие-то патенты не работают :)
- Создавайте хороший контент!

Вопросы?

Сергей Кокшаров

Devaka.ru