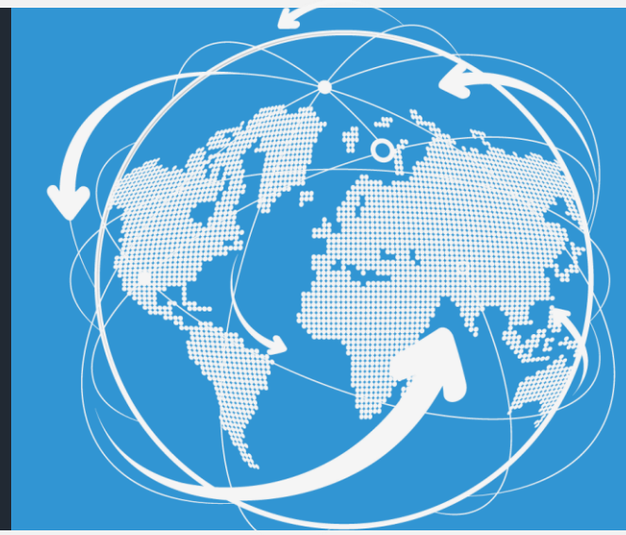
Weimob 微盟

国内最大的微信公众服务平台



全文检索

研发中心-基础平台　蔡林林

2016年2月26日

概要

- 全文检索
- 介绍 Apache Lucene
- 介绍 Elasticsearch
- 搜索服务

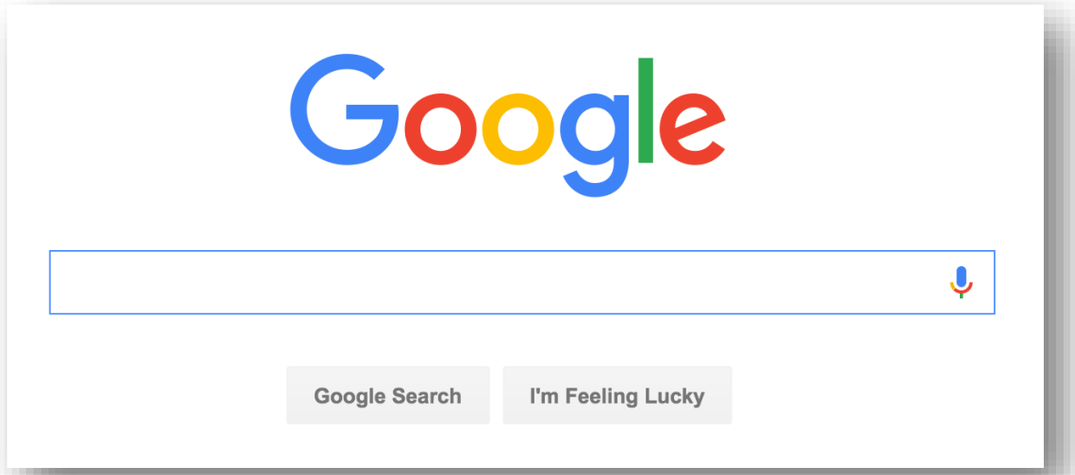
什么是搜索？

◆数据的种类：

- 结构化数据
- 非结构化数据（全文数据）

◆搜索的种类：

- 对结构化数据的搜索
- 对非结构化数据的搜索（全文检索）



全文检索(Full-text Search)是指以非结构化的纯文本信息作为检索对象的一种信息检索技术。

全文检索该怎么检索？

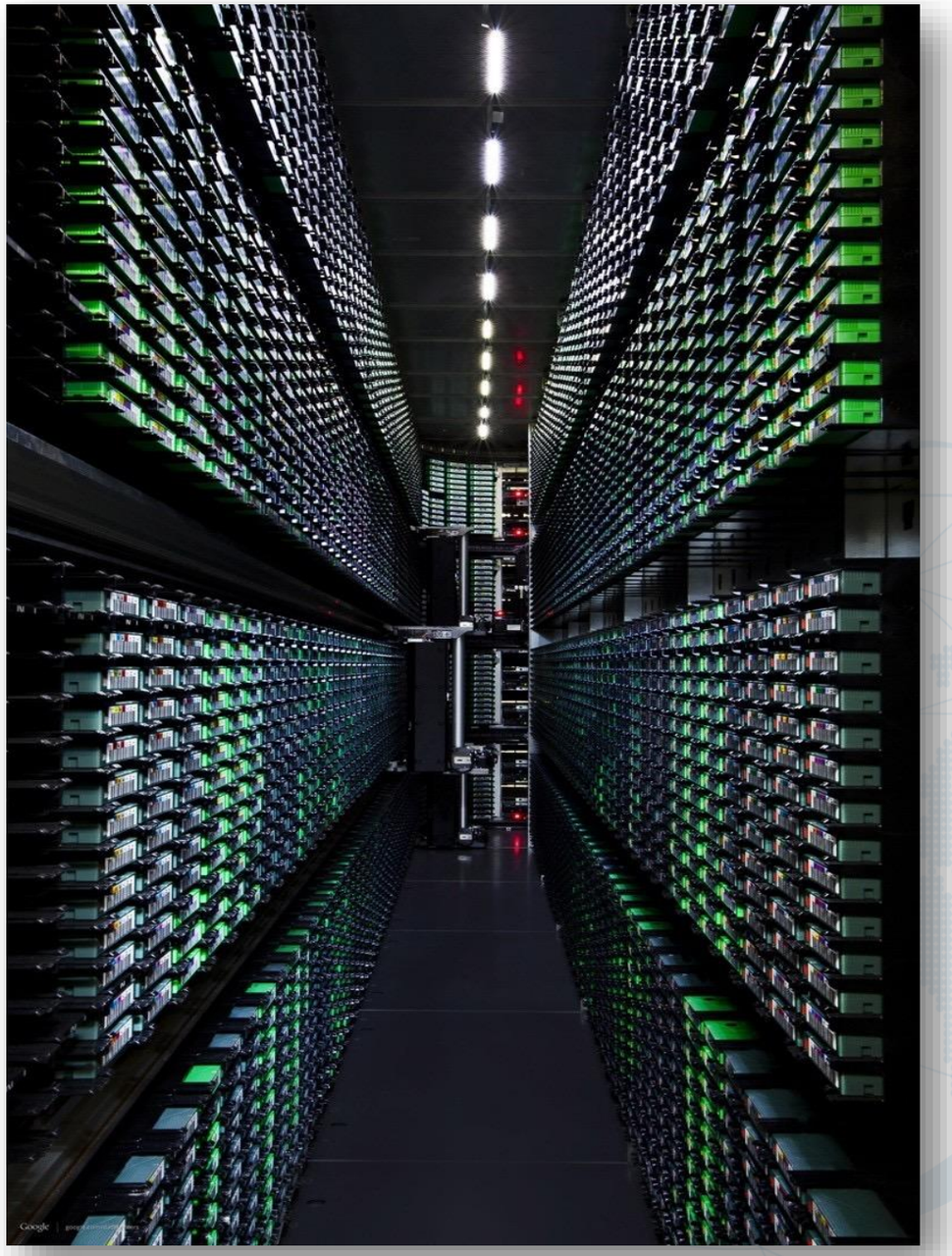
- 顺序扫描，字符串匹配
- 建立索引

为什么要用全文检索？

- 速度更快
- 结果排序

搜索要解决什么问题？

- 收集信息（如网络蜘蛛）
- 整理信息（文本分析生成索引）
- 接受查询（友好的查询方式）
- 结果排序（文本相关性等）

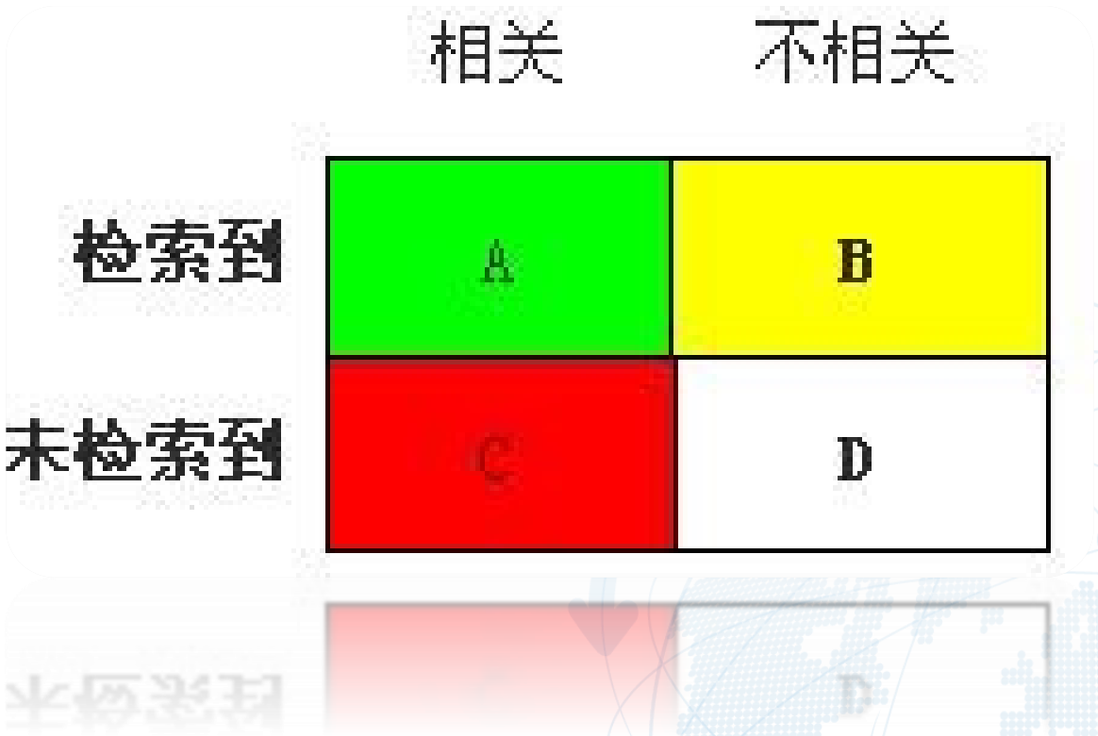


评价指标

- 召回率(Recall Rate)
- 准确率(Precision Rate)

$$R = A/(A+C)$$

$$P = A/(A+B)$$



全文检索

- 文本处理
- 倒排索引
- 文本相关性

文本处理

- 文本清理
- 文本分词

Doc: 习近平当选中华人民共和国主席

1: 习/近/平/当/选/中/华/人/民/共/和/国/主/席

2: 习/近/平/当选/中华/人民/共和国/主席

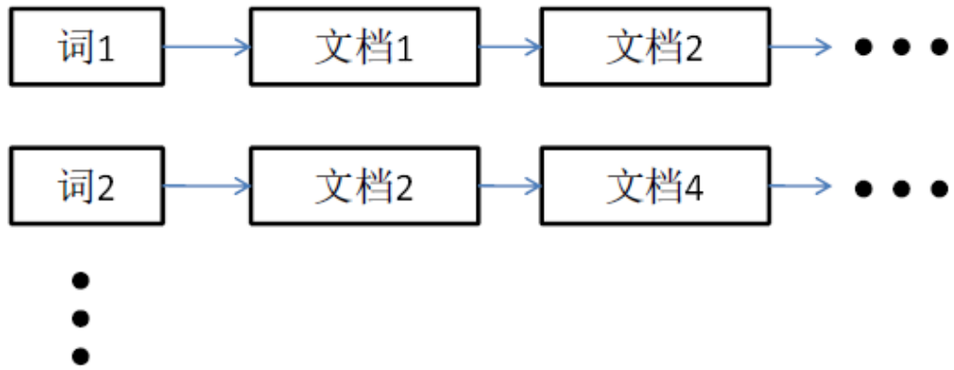
3: 习近平/当选/中华人民共和国/主席

4: 习近平/当选/中华人民共和国/中华/人民/共和国/主席

倒排索引（Inverted Index）

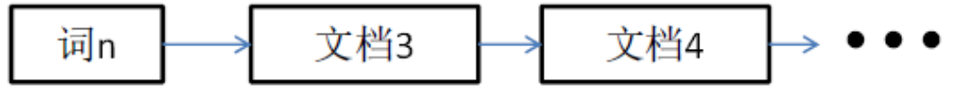
◆问题1:

这本书第 N 页有什么内容？



◆问题2:

这本书包含“中国”这个词的有哪些页？



倒排索引是一种数据结构，针对全文数据建立**关键词（term）**到**文档（document）**的关系，基于这种数据结构，有效提高全文检索的效率

文档1: 韩都衣舍韩版2016春装新款女百搭立领宽松显瘦牛仔短外套AA5494

文档2: 韩都衣舍2016春新款女印花图案显瘦长袖连衣裙秒杀

文档1: 韩都衣舍/韩版/2016/春装/新款/女/百搭/立领/宽松/显瘦/牛仔/短外套/AA5494

文档2: 韩都衣舍/韩版/2016/春/新款/女/印花/图案/显瘦/长袖/连衣裙/秒杀

Term	Docs
韩都衣舍	1，2
韩版	1，2
2016	1，2
春装	1
新款	1，2
女	1，2
百搭	1，2
立领	1
宽松	1
显瘦	1，2
牛仔	1
短外套	1
AA5494	1
春	2
印花	2
图案	2
长袖	2
连衣裙	2
秒杀	1

搜索“都衣”会出现什么结果？

搜索“HSTYLE”会出现什么结果？

文本相关性

1. 布尔模型

2. 空间向量模型

3. 概率模型

TF/IDF — 空间向量模型的一种实现

- ***N*，**集合中的文档总数。
- ***tf* ，**Term Frequency 的缩写。表示某个关键词在某个文档中出现的频率。
- ***df*，**Document Frequency 的缩写。表示文档集合中，出现某个关键词的文档个数。
- ***idf*，**Inverse Document Frequency 的缩写。表示 N 与 某关进词的 df 比值的对数。
- ***dl*，**Document Length 的缩写。表示文档长度。
- ***adl*，**Average Document Length 的缩写。表示平均文档长度。

$$\mathrm{tf}(t, d) = 0.5 + 0.5 \cdot \frac{f_{t,d}}{\max\{f_{t',d} : t' \in d\}}$$

$$\mathrm{idf}(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

$$\mathrm{tfidf}(t, d, D) = \mathrm{tf}(t, d) \cdot \mathrm{idf}(t, D)$$

基本原则

- 一个关键词在某个文档中出现次数越多（tf），这个关键词的权重越低。

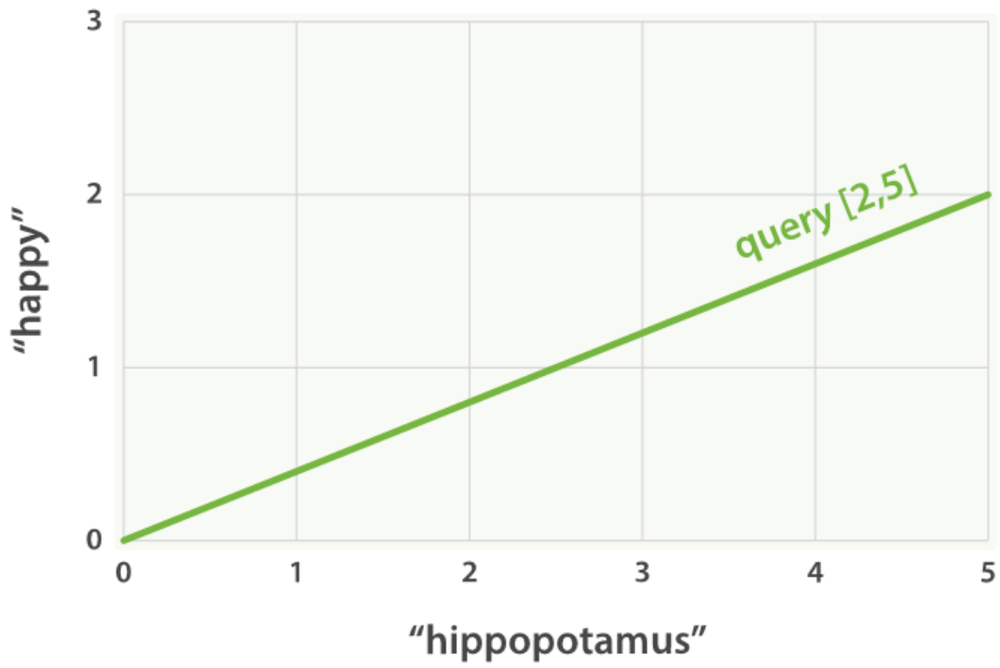
- 一个关键词在越多的文档中出现（df），这个词区分文档的作用就越低，这个关键词的权重也越低。

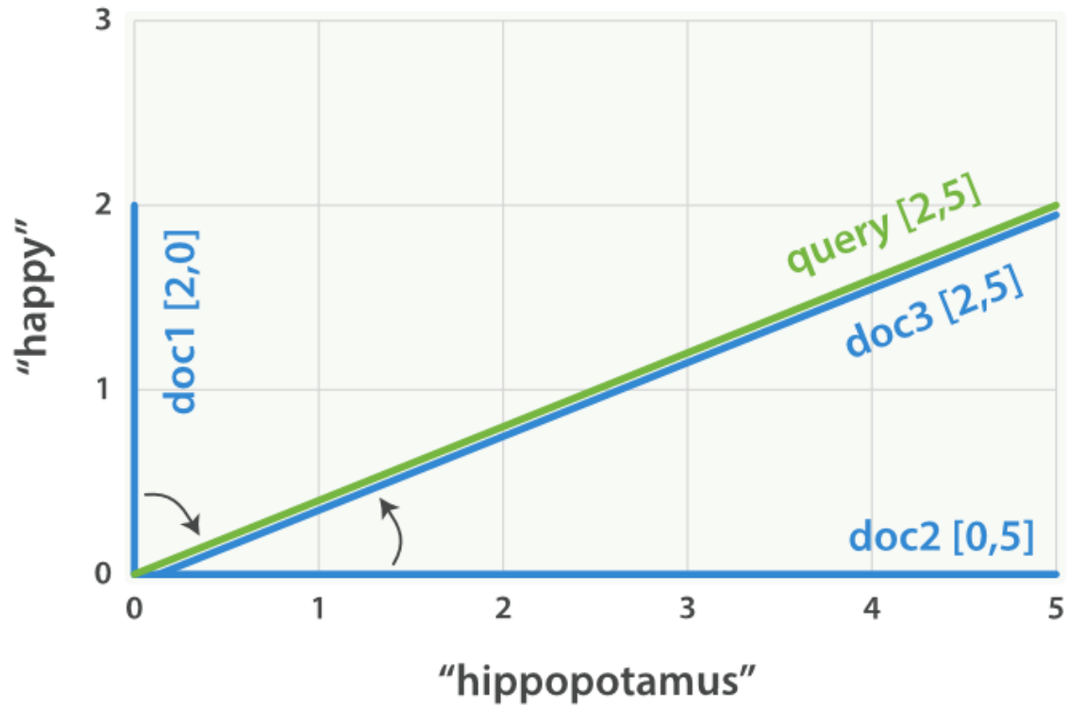
- 文档越长（dl），其出现某个关键词的次数可能越高，而每个关键词对这个文档的区分作用也越低，这些关键词的权重也越低。

1. I am **happy** in summer.
2. After Christmas I’ m a **hippopotamus**.
3. The **happy hippopotamus** helped Harry.

Query: “happy hippopotamus”

Term	weight
happy	2
hippopotamus	5





Apache Lucene

- **开源软件**
- **全文检索引擎框架**
  1、查询引擎
  2、索引引擎
  3、文本分析引擎
- **基于 Java 实现**

org.apache.Lucene.search	搜索入口
org.apache.Lucene.index	索引入口
org.apache.Lucene.analysis	语言分析器
org.apache.Lucene.queryParser	查询分析器
org.apache.Lucene.document	存储结构
org.apache.Lucene.store	底层IO/存储结构
org.apache.Lucene.util	一些公用的数据结构

Apache Lucene 中的索引

大部分搜索引擎（如数据库）采用的 B树 结构来维护索引，索引更新会导致大量的 IO 操作。Lucene 中的索引最大的特点是**不可变**（Immutable），每次索引文档都会创建一个新的 segment 。

- 控制磁盘 IO，提高索引效率
- 降低读写冲突，利于并发
- 方便数据扩展和恢复

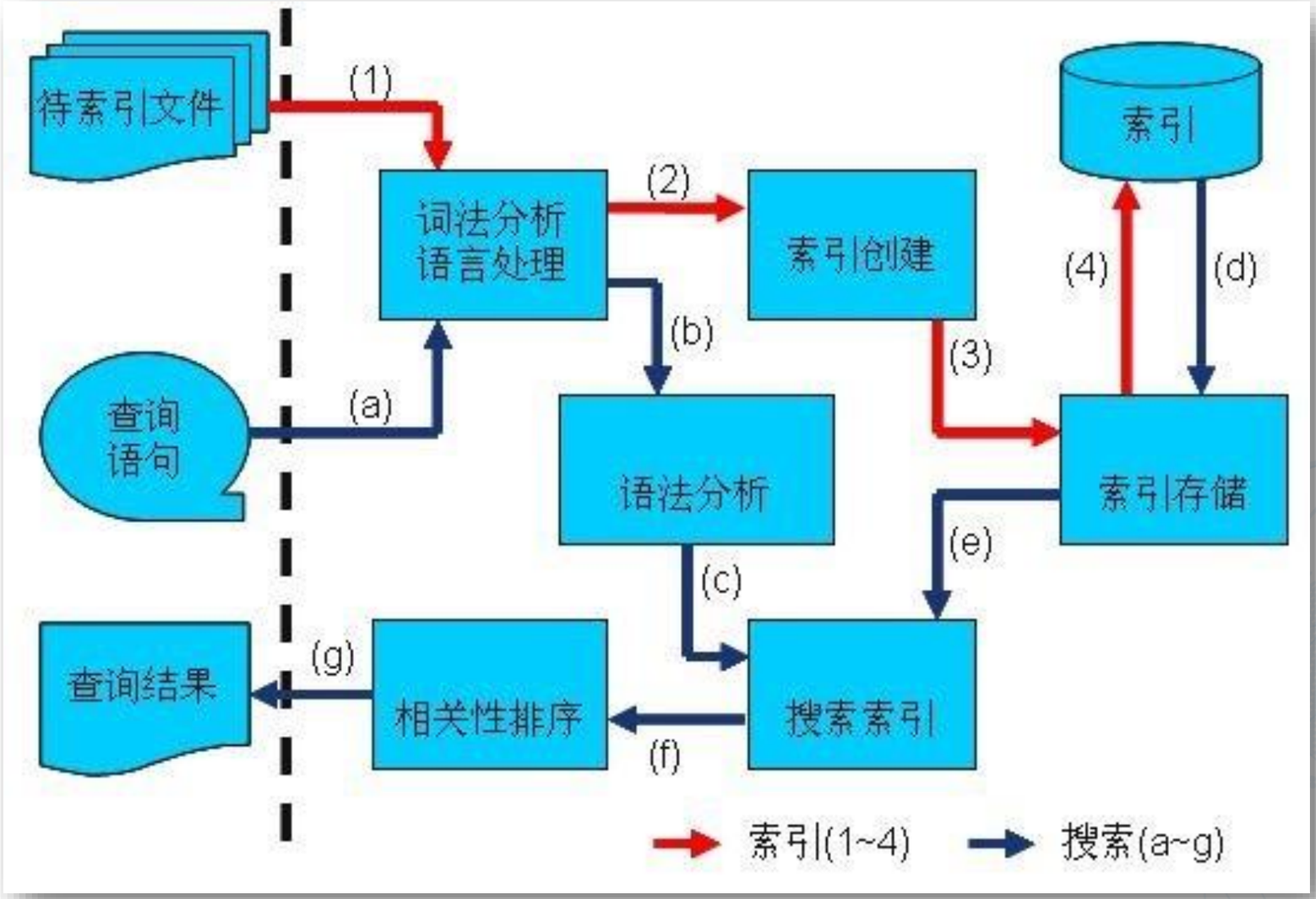
◆ 如何更新和删除文档？

不管是更新还是删除，都只在特定的文件（后缀为 .del）文件中记录某个文档是否删除。更新操作记录旧文档删除，更新后的文档保存在新的 segment，删除操作直接在记录文件中标记文档已删除。所以，整个索引过程是一个增量过程。在检索返回结果的时候根据 .del 文件清除不应该返回的文档。

◆ segment 越来越多怎么办？

定时的 merge segment。在保证机器资源的情况下，自动进行 segment 的 merge 操作，将小的 segment 合并成大的 segment。

Apache Lucene 工作流程

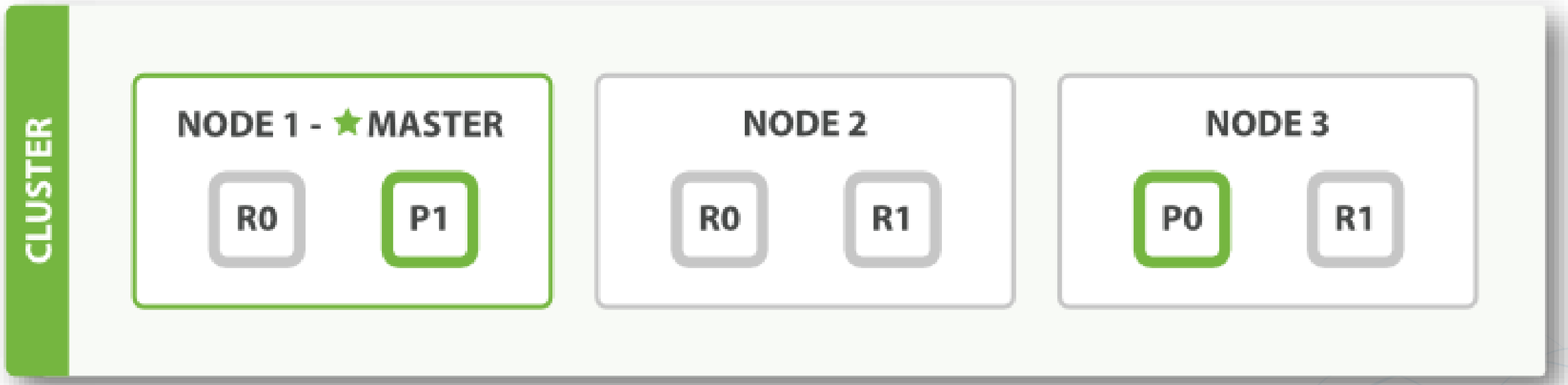


Elasticsearch

1. 简单方便的启动配置
2. 原生支持分布式模式
3. 对等网络架构，避免单点故障
4. 多节点配置，易于扩展
5. 方便的索引配置
6. 文档 versioning 机制，降低近实时搜索 Near Real Time (NRT)

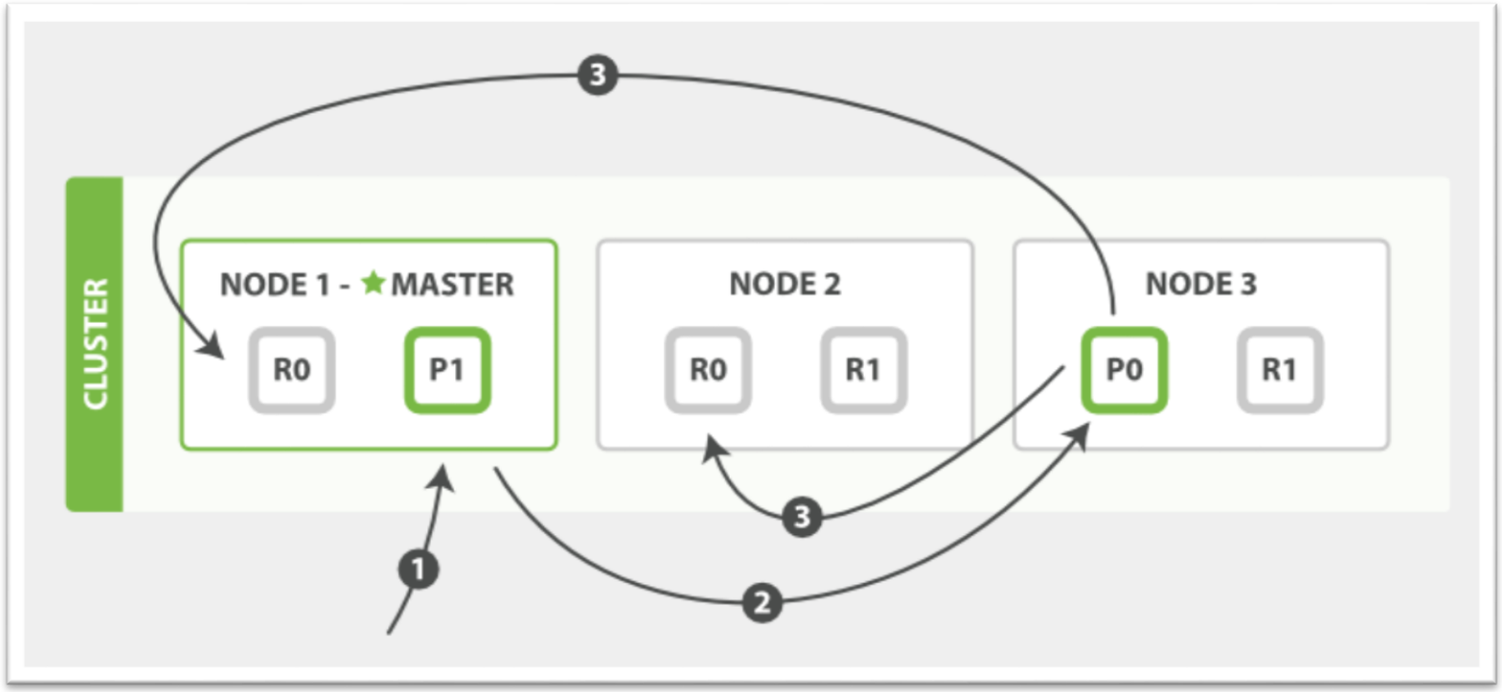
   的影响

- **Index**（索引）
- **Type**
- **Document 和 Field**
- **Mapping**
- **Cluster**（集群）
- **Node**（节点）
- **Primary Shard**（主分片）
- **Replica Shard**（副分片）



- 1 个集群
- 3 个节点，1个 Master node，2个 Data node
- 1 个索引
- 2 个主分片，4个副分片，1 个主分片对应 2 个副分片

Elasticsearch 分布式索引



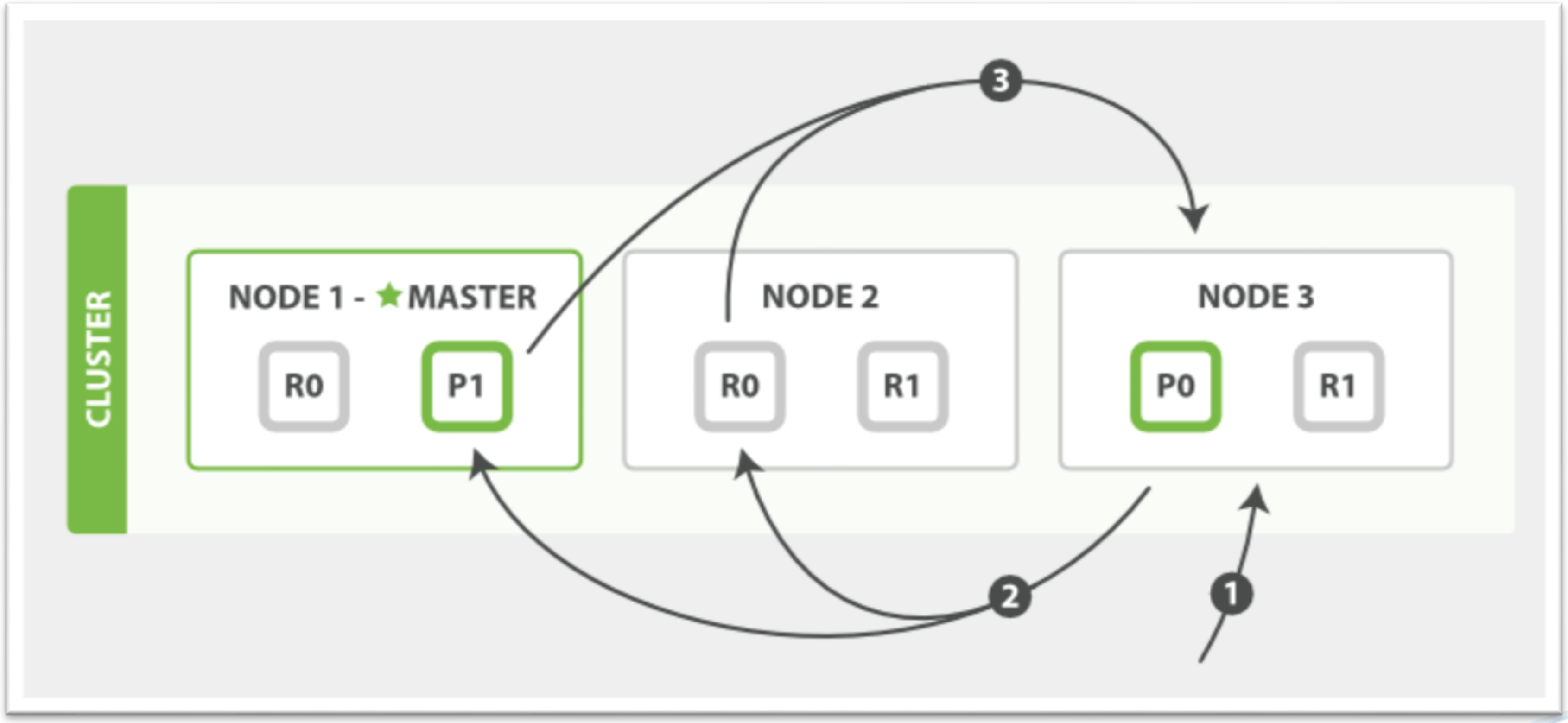
1. 节点1 收到索引请求
2. 节点1 根据_id决定索引在分片P0上
3. 主分片索引成功，推送到相应的副分片 R0

shard_num = hash(_routing) % num_primary_shards

Elasticsearch 分布式搜索

Query 查询结果

Fetch 获取详情



Query 过程

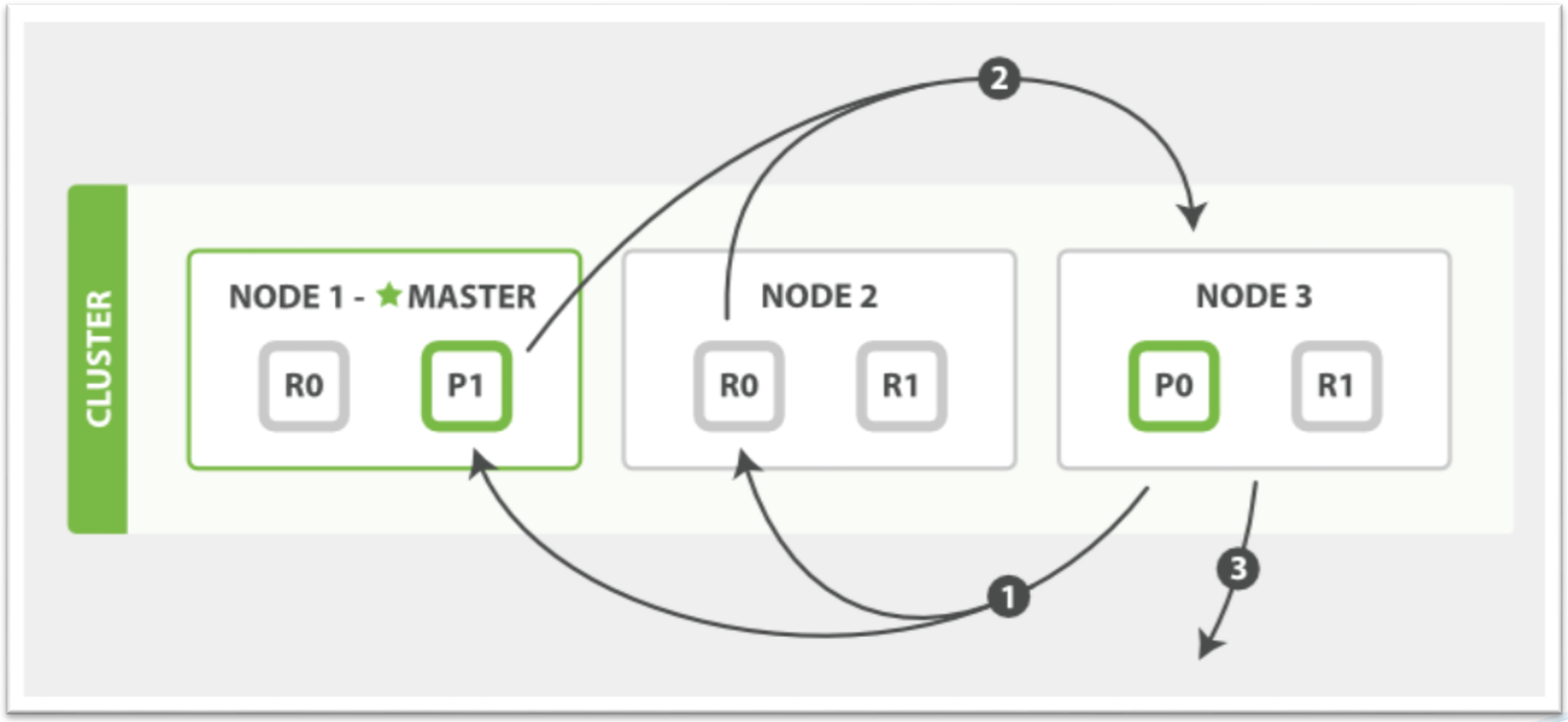
1. 节点3 收到检索请求，节点3 成为**协调节点**，并创建一个大小为 **from+size** 的优先队列

ES 的查询会要求 **from** 和 **size** 参数，来告知需要返回哪些结果，这就是分页

Query 过程中，从其他分片上返回的优先队列只包含文档的 _id 和 文档的 _score。

2. 节点3 将检索请求发送到节点1 上的 P1 和节点2 的 R0，分别创建大小为 **from+size** 的优先队列

3. 节点1 和 节点2 将结果返回到节点3，由节点3 整合到一个 大小为 **from+size** 的优先队列



Fetch 过程

1. 节点3 决定哪些文档需要取回，并根据 _id 分发到相应分片上。
2. 各分片取出文档，然后返回给节点3。
3. 节点3 收集和整理好所有文档，返回给客户端

翻页问题

Elasticsearch 进行深度翻页操作的代价是很昂贵的！

假设一个索引有 K 个主分片，一次的请求的 from 值为 F，size 值为 S。P 表示优先队列的大小。最终返回结果数为 S。

那么 P = F+S；协调节点要处理的结果数 N 为

$$N = K*P = K*(F+S)$$

Google 哈哈

All More Search tools

Sorry, Google does not serve more than 1000 results for any query. (You asked for results starting from 10000.)

Your search - 哈哈 - did not match any d ents.

Suggestions:

- Try different keywords.

Google

Previous

New York, NY - From your Internet address - Use precise location - Lea

Help Send feedback Privacy Terms

淘宝网 Taobao.com

宝贝 卫 搜索 在结果中排除 请输入要排除的词 确定

¥108.00 包邮 760人付款

韩国代购2015秋冬季新款韩版宽松长袖可爱上衣时尚休闲卡通卫衣女

0o波挂伊人牧o0 山东 青岛

¥136.00 包邮 35人付款

高端女装 2015秋冬新款韩国版宽松中长款连帽卫衣卡通米奇卫衣裙

派高格潮牌 广东 中山

¥ 52.00 包邮 56人付款

厨 浴室 洗手间 冲凉房 厨卫 卫浴 双层 物架 挂架 带钩壁挂

斯旗舰店 广东 中山

¥139.00 包邮 236人付款

冬季男士加绒加厚修身直筒运动裤卫裤羊羔绒收口卫裤休闲裤男女款

美丽_无极限 上海

潮流导购: 男士 韩版 套头 加绒 宽松 卫衣 学生 秋冬装 长袖 休闲

< 上一页 1 2 3 4 5 ... 下一页 > 共 100 页，到第 2 页 确定

全文检索的缺陷

返回很多与 搜索意图 不相关的结果

改进方式　提高精确度和召回率

- Keywords
- Boolean queries
- Phrase search
- Proximity search
- Regular expression
- Wildcard search

改进方式

- 分类器，提升内容相关性
- 增加排序因子，改进排序模型

垂直化和精细化的搜索服务

- 搜索与场景相关（时间、地点）
- 搜索与内容相关（新闻、网页、商品）
- 搜索与用户相关（小白用户、高级用户）
- 查询纠正和效果反馈
- 数据更新效率和请求响应速度

Bai du 百度

上海明天天气怎么样

百度一下

网页 新闻 贴吧 知道 音乐 图片 视频 地图 文库 更多»

百度为您找到相关结果约3,580,000个

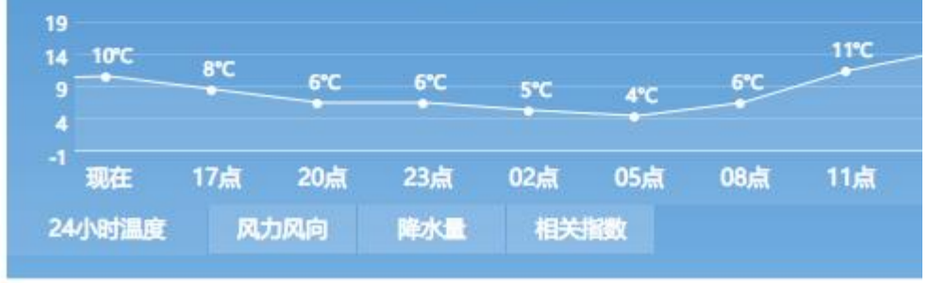
搜索工具

上海天气预报_一周天气预报_中国天气网 - 最近访问：北京天气

周四 02月25日 农历正月十八 (实时：10℃)

11 ~ 4℃
晴转多云
东南风微风
实时空气质量: 86 良

周五 02月26日	周六 02月27日	周日 02月28日
15 ~ 7℃	18 ~ 9℃	17 ~ 5℃
多云	晴	阴
南风微风	南风微风	南风微风



中国气象局2016年02月25日15时发布 7天预报 8-15天预报 周边景点天气

Google

上海明天天气怎么样

All Maps News Images Videos More Search tools

About 649,000 results (0.44 seconds)

Shanghai, China
Fri
Partly Cloudy

16 °F | °C

Precipitation: 0%
Humidity: 54%
Wind: 19 km/h

Temperature Precipitation Wind

Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu
9° 4°	16° 6°	19° 6°	19° 2°	9° 0°	11° 3°	16° 5°	17° 9°

More on weather.com

Feedback

Bai du 百度

上海后天天气怎么样

百度一下

网页 新闻 贴吧 知道 音乐 图片 视频 地图 文库 更多»

百度为您找到相关结果约1,620,000个

搜索工具

上海天气预报_一周天气预报_中国天气网 - 最近访问：北京天气

中国气象局2016年02月25日15时发布 7天预报 8-15天预报 周边景点天气

Google

上海后天天气怎么样

All Maps Images News Videos More Search tools

About 736,000 results (0.42 seconds)

Shanghai, China
Sat
Sunny

19 °F | °C

Precipitation: 0%
Humidity: 49%
Wind: 16 km/h

Temperature Precipitation Wind

Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu
9° 4°	16° 6°	19° 6°	19° 2°	9° 0°	11° 3°	16° 5°	17° 9°

More on weather.com

Feedback

Bai du 百度

上海下周一天气怎么样

百度一下

网页 新闻 贴吧 知道 音乐 图片 视频 地图 文库 更多»

百度为您找到相关结果约4,790,000个

搜索工具

【上海天气】上海天气预报,上海天气预报一周,15天天气查询-2345...

天气条件较不宜晨练。 紫外线中等 外出需要涂抹中倍数防晒霜 易发感冒 体质较弱的朋友需注意防护。 较适宜晾晒 请在室外地方晾晒。 适...

tianqi.2345.com/shangh... - 百度快照 - 73%好评

【上海天气预报】上海天气预报一周_上海天气预报10天、15天查询—

天气网 www.tianqi.com

★天气网(www.tianqi.com)★上海天气预报提供上海今日天气、气以及上海未来一周的天气预报,可以实时查看上海天气预报一天、15天的天气情况。旅游出行,...

shanghai.tianqi.com/ - 百度快照 - 78%好评

上海天气预报_上海天气预报一周7天10天15天_上海市未来一周天气 -

上海天气预报查询;中央气象台发布的最新上海天气预报一周7;15天查询、上海市未来一周天气查询、气温状况风力等气象查上海各辖区的天气数据信息查询!

15tianqi.cn/shanghai/ - 百度快照 - 91%好评

Google

上海下周一天气怎么样

All Maps News Images Videos More Search tools

About 554,000 results (0.21 seconds)

Shanghai, China

Mon

Sunny

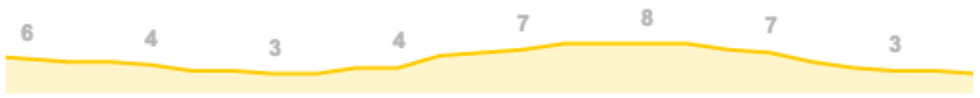
9 °C | °F

Precipitation: 0%

Humidity: 41%

Wind: 23 km/h

Temperature Precipitation Wind



Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu
9° 4°	16° 6°	19° 6°	19° 2°	9° 0°	11° 3°	16° 5°	17° 9°

More on weather.com

Feedback

主要参考资料：

- Rafal Kuc, Marek Rogozinski . Mastering ElasticSearch. Packt Publishing ,2013.
- Clinton Gormley, Zachary Tong . Elasticsearch: The Definitive Guide.O'Reilly Media,2015.
- Lucene_3.0_原理与代码分析完整版,

http://www.cnblogs.com/forfuture1978/archive/2009/12/14/1623594.html

Weimob 微盟

国内最大的微信公众服务平台

thank you !