Finding Similar Sets

Applications
Shingling
Minhashing
Locality-Sensitive Hashing

Mining of Massive Datasets
Leskovec, Rajaraman, and Ullman
Stanford University

Applications of Set-Similarity

Many data-mining problems can be expressed as finding "similar" sets:

1. Pages with similar words, e.g., for classification by topic.
2. NetFlix users with similar tastes in movies, for recommendation systems.
3. Dual: movies with similar sets of fans.
4. Entity resolution.

定义：不同的数据提供方对同一个事物即实体 (Entity)可能会有不同的描述 (这 里的描述包括数据格式 、表示方法 等) ，每一个对实体的描述称为该实体的一个引用。实体解析，是指从一个“ 引用集合 ”中解析并映射到现实世界中的“ 实体 ”过程 。

实体解析(Entity Resolution)又被称为记录链接(Record Linkage) 、对象识别(object Identification) 、个体识别(Individual Identification) 、重复检测(Duplicate Detection)
https://www.cnblogs.com/nolonely/p/5399695.html

Similar Documents

- Given a body of documents, e.g., the Web, find pairs of documents with a lot of text in common, such as:
  - Mirror sites, or approximate mirrors.
    - Application: Don't want to show both in a search.
  - Plagiarism, including large quotations.
  - Similar news articles at many news sites.
    - Application: Cluster articles by "same story."

核心思想就是从一些修改或者被污染的数据中找出来和原来的东西一样的或者是镜像：比如抄袭，或者是相同类容的新闻或者文章筛选

Three Essential Techniques for Similar Documents

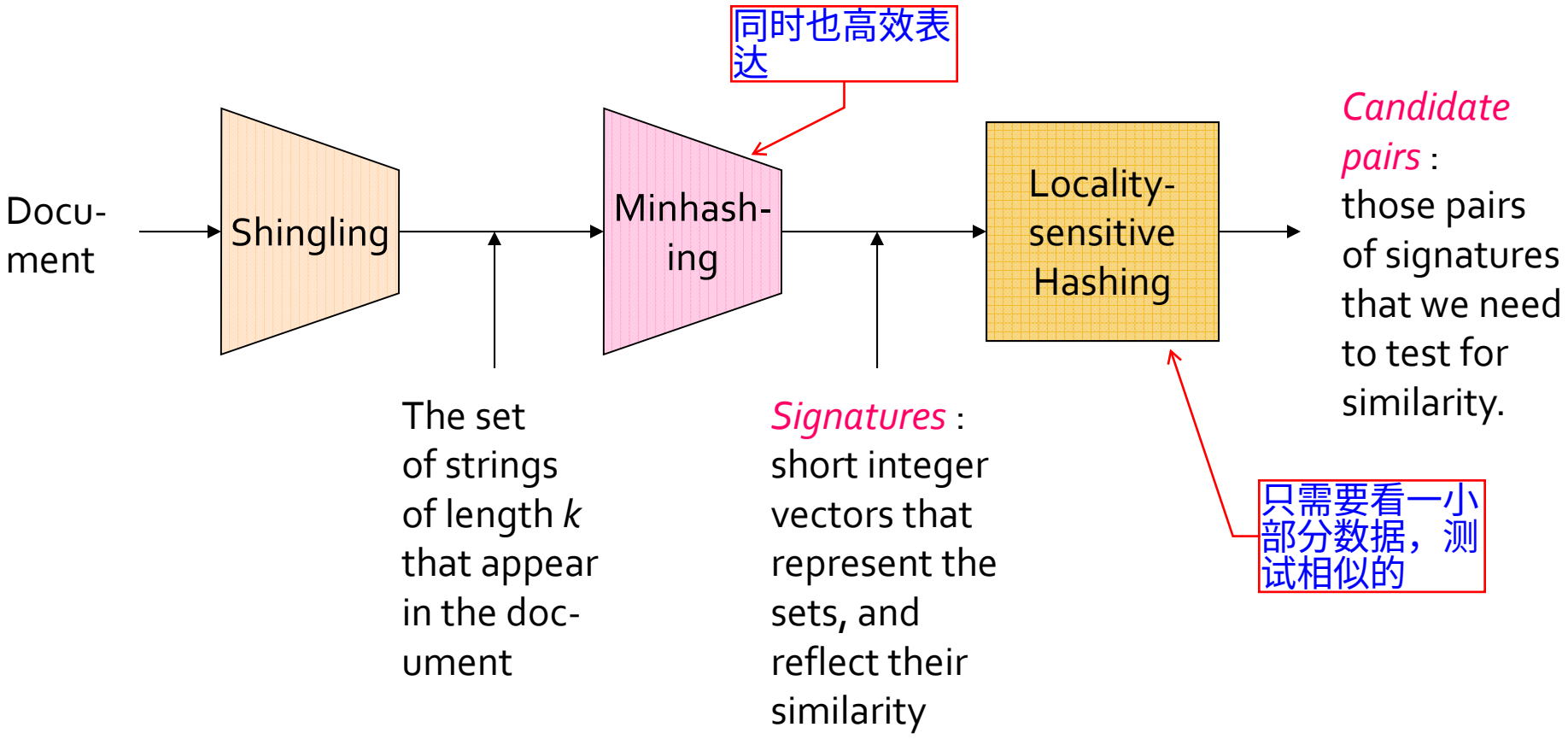
还可以告诉我们，shingling搞出来的东西究竟有多相似

1. *Shingling* : convert documents, emails, etc., to sets.
2. *Minhashing* : convert large sets to short signatures, while preserving similarity.
3. *Locality-sensitive hashing* : focus on pairs of signatures likely to be similar.

a lot of text in common

LSH最根本的作用，就是能高效处理海量高维数据的最近邻问题
Locality-sensitive hashing (LSH) reduces the dimensionality of high-dimensional data. LSH hashes input items so that similar items map to the same “buckets” with high probability (the number of buckets being much smaller than the universe of possible input items).
Locality-sensitive hashing has much in common with data clustering and nearest neighbor search.

The Big Picture



Shingles

- A *k* -shingle (or *k* -gram) for a document is a sequence of *k* characters that appears in the document. k=5,10经常用
- Example: k=2; doc = abcab. Set of 2-shingles = {ab, bc, ca}.
- Represent a doc by its set of *k*-shingles.

Shingles and Similarity

- Documents that are intuitively similar will have many shingles in common.
- Changing a word only affects k-shingles within distance k from the word.
- Reordering paragraphs only affects the 2k shingles that cross paragraph boundaries.
- Example: k=3, “The dog which chased the cat” versus “The dog that chased the cat”.
  - Only 3-shingles replaced are g_w, _wh, whi, hic, ich, ch_, and h_c.

7个shingles

6个shingles

Shingles: Compression Option

http://ethen8181.github.io/machine-learning/clustering_old/text_similarity/text_similarity.html

- To compress long shingles, we can hash them to (say) 4 bytes.
  - Called *tokens*.
- Represent a doc by its tokens, that is, the set of hash values of its *k*-shingles.
- Two documents could (rarely) appear to have shingles in common, when in fact only the hash-values were shared.