李胜利

北京大学信息管理系

副教授

嘉宾简介

- 博士毕业于美国佛罗里达大学
- 现任北京大学信息管理系副教授

DQMiS

第二届数据质量管理国际峰会

The 2nd Data Quality Management International summit

北京大学
PEKING UNIVERSITY

在线医疗信息质量——基于搜索引擎的研究

主讲人：李胜利

2018年9月

目录

CONTENTS

01

研究背景

研究背景

通过互联网进行医疗信息的搜索正在变得日益流行

on any given day, more people are posing health questions to Google than posing health questions to their doctors

61%的美国成年人通过网络获取医疗信息

其中65%使用搜索引擎

搜索引擎对于用户在医疗方面的决策有很大影响

研究背景

- 46%的用户在完成搜索以后，并没有去找医生做进一步的咨询
- 低质量的搜索结果有可能误导用户，延误治疗时机

研究背景

魏则西事件

- 魏则西大二时发现患上滑膜肉瘤，通过百度搜索，他到武警北京市总队第二医院（简称“武警二院”）尝试“肿瘤生物免疫疗法”，然而这种疗法在美国早已被淘汰，与斯坦福大学的合作也是虚假宣传，魏则西及网友质疑武警二院及相关医生存在欺骗行为；因医学信息竞价排名而饱受争议的百度也再次被质疑。
- 魏则西16年已经去世

研究背景

DQMiS

研究目标

搜索引擎是否把高质量的结果放到前面，低质量的结果放到后面。

几个因素可能会有负面的影响

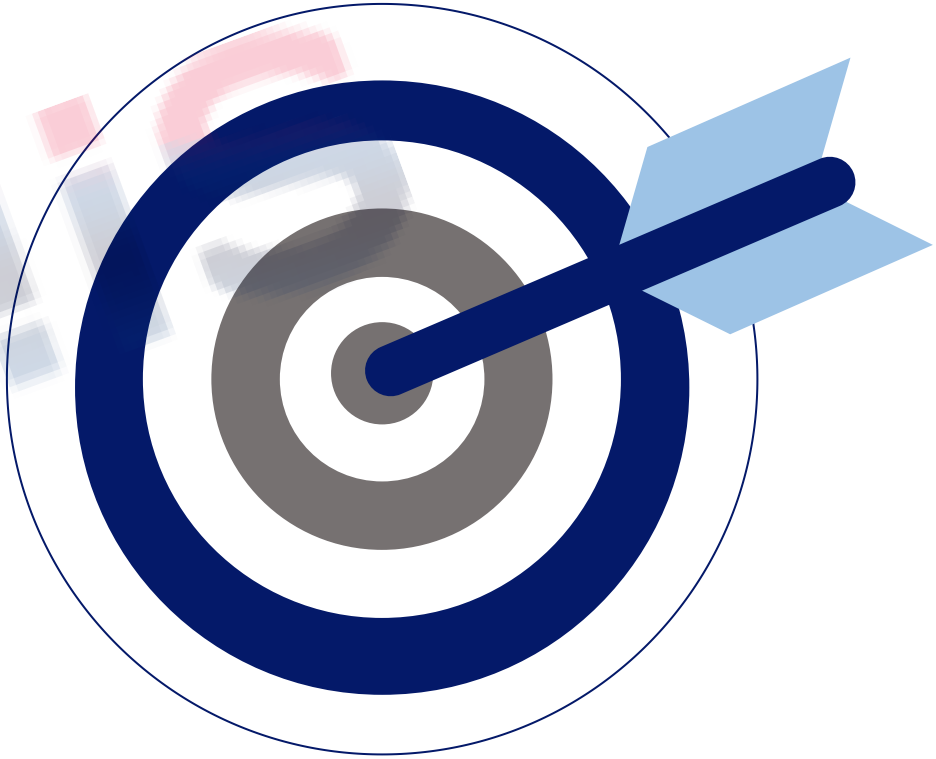
- 大量的用户搜索并使用一些网站，例如一些论坛等，会导致低质量结果排名高
- 低质量的网站通过SEO技术提高自己的排名
- 竞价排名的影响

02

研究方法

研究方法

- **收集了National Library of Medicine’s MedlinePlus 所包含的2069个医疗名词**
  - 这些名词接近用户搜索使用的关键词，包括“abdominal pain（腹痛）”等
- **对每一个名词，我们进行一次Google搜索**
  - 只取第一页的返回结果，每页包括十个搜索结果
  - 只保留Organic结果，去掉Sponsored结果（因为付费广告是有标识的）



研究方法

网站的评价

是否被HON Foundation (Health on the Net)或者MedlinePlus所收录？

2069个名词共返回了5249个网站

根据收录情况，返回的网站共分为几种情况：certified (MedlinePlus), certified (HON), referenced, not certified/referenced

我们可以定义certified或者reference都是高质量网站，也可以定义只有certified才是高质量网站

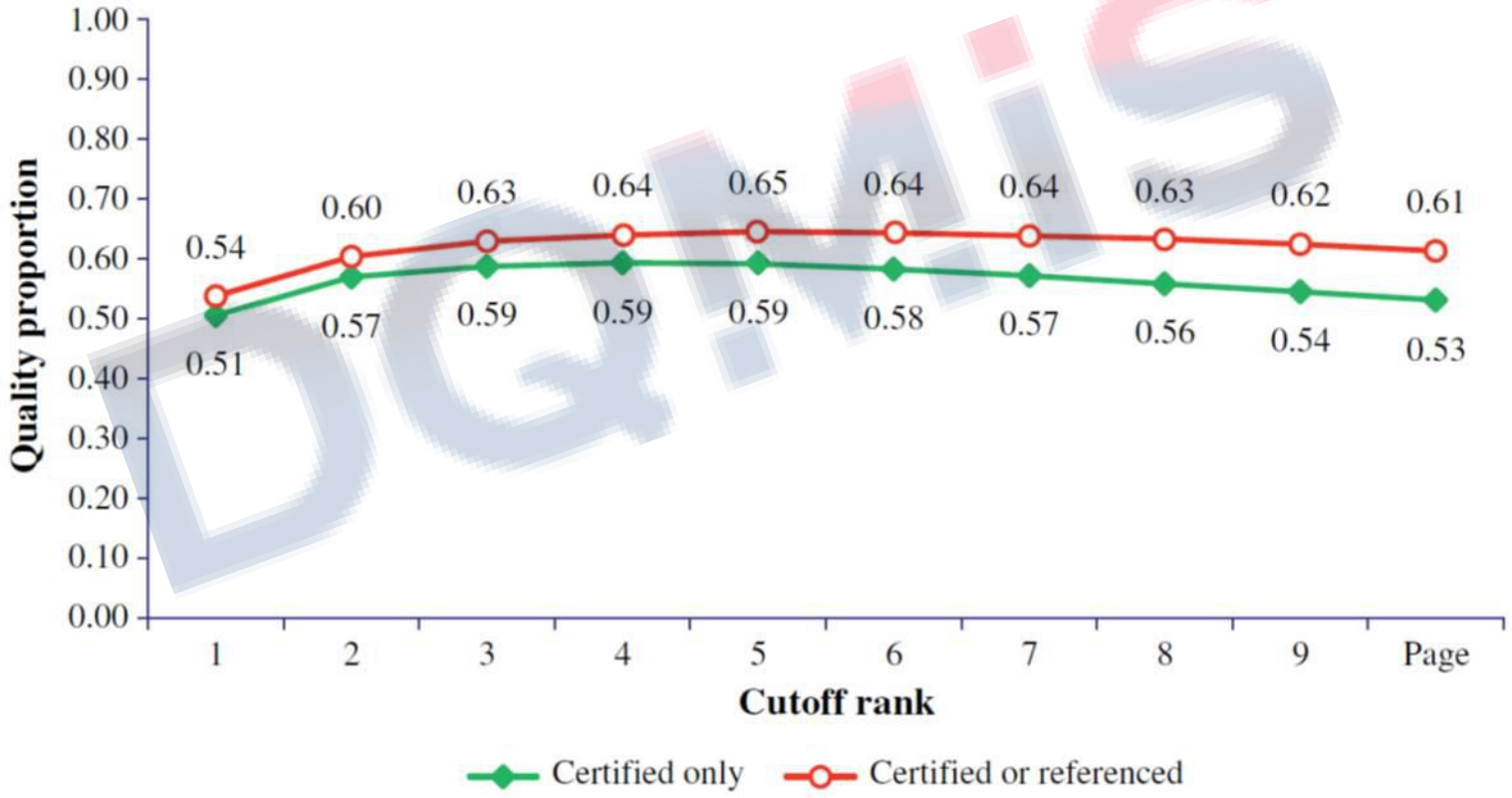
这样的分类方法在医疗领域是被广泛认可的

03

研究结果

研究结果

整体的信息质量



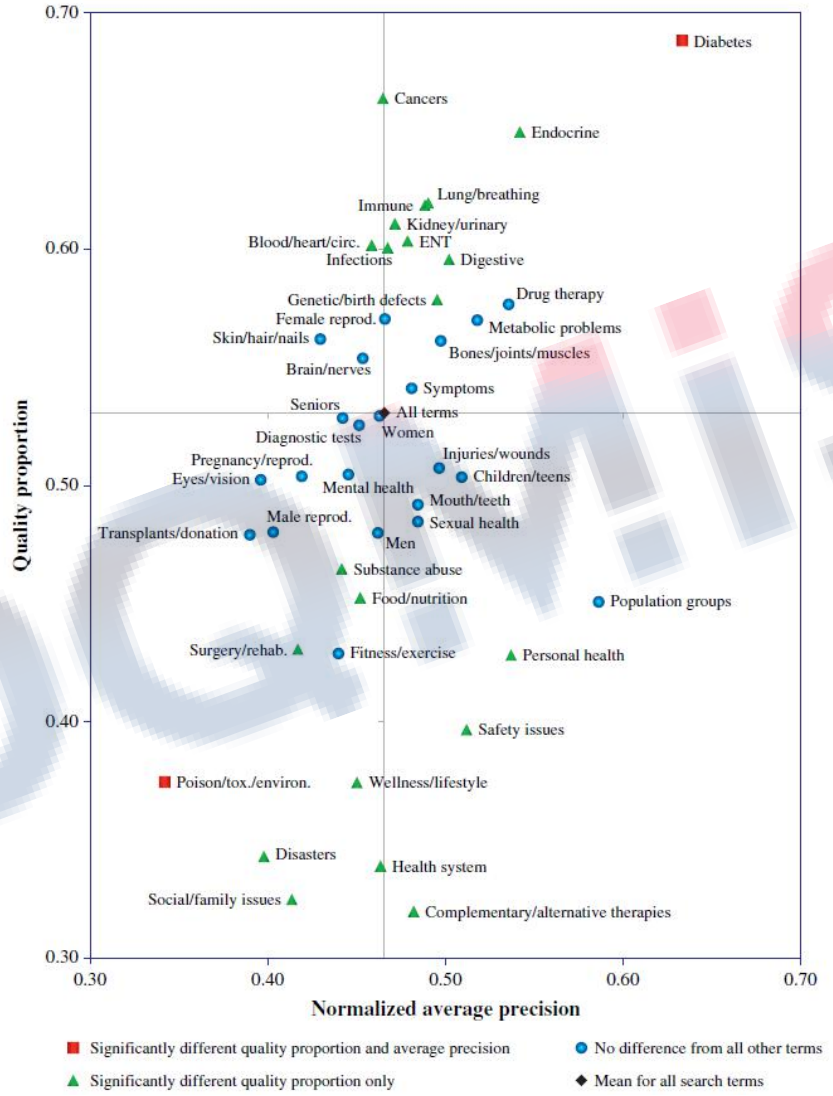
DQMiS

分类别信息质量评价

通过quality proportion与average precision (平均正确率)两个指标来衡量

$$\text{Average precision} = \frac{\sum_{i=1}^{|D|}\left(r(d_i)\left(\frac{\sum_{j=1}^{i} r(d_j)}{i}\right)\right)}{|D|}$$

研究结果


0.70
0.60
0.50
0.40
0.30
Quality proportion
0.30
0.40
0.50
0.60
0.70
Normalized average precision
Diabetes
Cancers
Endocrine
Lung/breathing
Immune
Kidney/urinary
Blood/heart/circ.
ENT
Infections
Digestive
Genetic/birth defects
Drug therapy
Female reprod.
Metabolic problems
Skin/hair/nails
Bones/joints/muscles
Brain/nerves
Symptoms
Seniors
All terms
Women
Diagnostic tests
Injuries/wounds
Pregnancy/reprod.
Children/teens
Eyes/vision
Mental health
Mouth/teeth
Male reprod.
Sexual health
Transplants/donation
Men
Substance abuse
Food/nutrition
Population groups
Surgery/rehab.
Fitness/exercise
Personal health
Safety issues
Poison/tox./environ.
Wellness/lifestyle
Disasters
Health system
Social/family issues
Complementary/alternative therapies
Significantly different quality proportion and average precision
No difference from all other terms
Significantly different quality proportion only
Mean for all search terms

04

研究启示

研究启示

整体的信息质量是不错的

需要搜索引擎特别注意的是，在某些类别下，返回信息质量不高，尤其是Social和preventive类别

DQMiS

第二届数据质量管理国际峰会

The 2nd Data Quality Management International summit

北京大学 PEKING UNIVERSITY

Thank You!