大数据典型应用调研报告

王艺羲 211250175

2023 年 9 月 9 日

1 调研

关于 Google 对 BDA 实际应用的报告

1.1 独立找到一个 BDA 在各个行业中的实际应用

第一个就是 ReCAPTCHA 案例，这个虽然是被谷歌收购的，但是，具有典型的谷歌思维。

为了解决垃圾邮件和网络机器人的问题，冯. 安发明了验证码的解决方案。如果只限于此，也就没有特别可以称道的，但是他意识到每天有这么多人要浪费 10 秒钟的时间输入这堆恼人的字母，而随后大量的信息被随意地丢弃时，他开始寻找能使人的计算能力得到更有效利用的方法。

他想到了一个继任者，恰如其分地将其命名为 ReCaptcha。和原有随机字母输入不同，人们需要从计算机光学字符识别程序无法识别的文本扫描项目中读出两个单词并输入。其中一个单词其他用户也识别过，从而可以从该用户的输入中判断注册者是人; 另一个单词则是有待辨识和解疑的新词。

为了保证准确度，系统会将同一个模糊单词发给五个不同的人，直到他们都输入正确后才确定这个单词是对的。在这里，数据的主要用途是证明用户是人，但它也有第二个目的：破译数字化文本中不清楚的单词。ReCaptcha 的作用得到了认可，2009 年谷歌收购了冯 · 安的公司，并将这一技术用于图书扫描项目，再后来，谷歌街景也开始使用这项技术。

把验证码和 OCR 需求巧妙结合起来，这展示了思维的威力，实现了 ReCaptcha 技术提供者和使用者的双赢，技术提供者利用 OCR 识别获得了自己的受益，使用者不需要任何付费 (互联网免费思维)，也愿意使用，对于用户其实也没有影响，没有增加额外的工作。上研究生的时候，就研究

OCR 汉字识别问题，识别率始终是个问题，对于手写就更低了，要花费大量人力来解决，并且，人工识别工作是非常无聊，没有办法来保障质量。再想起 12306 的验证码，更令人无语了。我们浪费了多少资源? 我们有多少资源可用充分来利用?

第二个是拼写检查纠错的案例。

我们都经常使用微软的 Word，其中就有拼写检查纠错功能，微软实现这个功能，采用的是传统的软件思维，也就是利用规则和词库来解决，这个需要不断耗费人力进行规则和词库的升级，对于不同的语言，耗费更是巨大。

谷歌解决这个方法，用的相对巧妙，在搜索的时候，当你输入一个错误的词时，会给一个提示，要找的是不是建议的词，如果用户确认后，谷歌就进行记录处理，后面，再经过一些算法处理，经过大量的数据学习，各个拼写检查纠错就越来越好，并且，这个后续维护成本很低，效果越来越好。

其实，谷歌翻译也使用了类似的思路，虽然前期算法，包括大数据处理花费了比较多，后续，基本实现了自动化，系统会越来越强，维护升级成本很低，项目就变成可持续发展。

2 报告

2.1 所属行业

谷歌是一家科技公司，主要在互联网和软件开发领域活动。它是科技行业的重要参与者，致力于提供各种在线服务和工具，包括搜索引擎、云计算、广告、地图和手机操作系统等。

2.2 解决的特定问题

在科技行业，谷歌使用 BDA 来解决多个特定问题，其中之一是搜索结果的改进。谷歌搜索引擎每天处理数十亿次搜索请求，用户期望获得与其查询最相关的结果。因此，谷歌需要不断改进其搜索算法以提供更准确的搜索结果，以满足用户需求。

2.3 使用的数据类型

为了改进搜索结果，谷歌使用了多种类型的数据，包括用户搜索历史、点击数据、网页内容、地理位置数据等。这些数据来自于各种来源，包括谷歌搜索引擎、谷歌地图、谷歌广告平台等。

2.4 BDA 如何有助于解决问题

BDA 在谷歌的应用中起到了关键作用。谷歌通过分析大规模数据集，利用机器学习算法和人工智能技术来改进搜索结果的质量。通过分析用户搜索历史，谷歌可以了解用户的兴趣和倾向，从而提供更加个性化的搜索结果。点击数据和网页内容分析有助于识别高质量的网页，以提高搜索结果的准确性。地理位置数据则可以用于提供本地化的搜索结果。总之，BDA 帮助谷歌不断优化搜索引擎，提供更好的用户体验。

2.5 BDA 对该行业未来可能的影响

BDA 对科技行业的未来影响巨大。随着数据规模的不断增长，谷歌将能够更好地理解用户需求，提供更个性化的服务。此外，BDA 还可以用于改进广告定位和效果分析，从而提高广告收入。随着人工智能技术的发展，谷歌还可以探索更多创新的应用，如自动驾驶汽车和智能家居。

3 结论

BDA 在科技行业中的应用，如谷歌的搜索引擎优化，展示了其在实际问题解决中的巨大潜力。通过分析多种数据类型，BDA 可以帮助企业更好地理解用户需求，提供更个性化的服务，从而在竞争激烈的科技行业中取得竞争优势。未来，BDA 将继续在科技行业中发挥重要作用，推动行业的创新和发展。