河 海 大 學

《信息组织检索》实验报告

成员姓名 杨中昊、周承桂、何雷、漆银权、李宗霖

专业年级 20 级大数据管理与应用

学年学期 2022-2023 学年 1

指导老师 冯兰萍

2022 年 11 月

实验一

1.检索策略设计

1.1 实验检索课题

新时代高校大学生就业质量影响因素分析

研究问题	新时代高校大学生就业质量影响因素分析
研究目标	形成文献综述，大学生就业质量影响因素分析
时间范围	10 年
地域范围	全国
背景/事件	高校毕业生数量不断增加，就业形势严峻

1.2 提取检索词

中文：新时代、高校大学生，同义词高校毕业生、就业质量

英文：New Era，College students，College graduates，Employment quality

1.3 构建检索表达式

中文：新时代 AND （高校大学生 OR 高校毕业生） AND 就业质量

英文：New Era AND （College students OR College graduates） AND Employment quality

1.4 检索信息源

1.4.1 信息资源类型

经过小组成员讨论，最终选择使用度较高的 CNKI 作为本次实验的检索系统，同时选择知网内的文献中的标题、摘要、关键词作为语料分析的对象，用于后续的相关评价。

1.4.2 CNKI 提供的检索途径

(1) 简单检索

即为在知网主页中直接采用文献检索方式，该方式提供了主题，篇关键，关键词，篇名，全文，作者，分类号，基金等途径，用户可以基于自己已有的信息内容，选择对应的文献形式进行检索。如下图所示：

图 1-1 简单检索界面

(2) 知识元检索

基于用户输入的问题，选择需要获取的形式，比如知识问答、百科或者词典等等来进行检索。

图 1-2 知识元检索页面

(3) 引文检索

基于用户检索文献中被引用的特征词进行的检索方式，特征词有被引主题、被引题名、被引关键词、被引摘要、被引作者、被引单位、被引文献来源。

(4) 高级检索

即为将简单检索的多个特征词进行组合进行精确或者模糊匹配的检索，基于用户需要灵活选取需要结合的特征词，如下所示：

图 1-3 高级检索页面

(5) 专业检索

使用检索语言,即符号运算符、逻辑运算符、括号运算符和检索项相结合构造检索表达式,达到快速搜索、精确定位的目的。

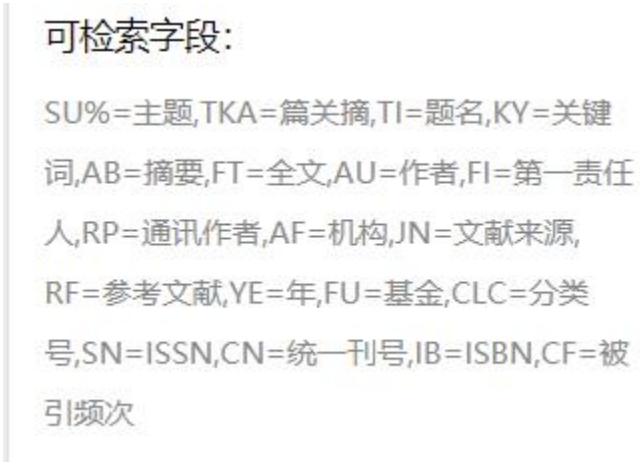

图 1-4 专业检索页面

除此之外还有按照作者发文检索以及句子检索，检索方式即为字面表述形式进行。

1.5 构建查询表达式

① 新时代 and 高校大学生 and 就业质量

② 新时代 and 高校毕业生 and 就业质量

2.信息组织方法分析

2.1 分类组织法

采用分类组织法，得到的类目体系划分结构较为直观，对于分类涉及的范围也是能够清晰的展现。用户可以按照划分的各个类型进行文献的筛选，一定程度上也可提高文献检索的效率。

在知网中，该方法被广泛使用。比如在首页面最顶端的一行便确定了检索范围，按照文献的数据库来源，对文献进行二级划分，各个具体的检索范围即为二级类目；但是所有的类目之间并不是严格意义上有区分，也就是对于同一文献可能存在于多个类目之下，在该体系中并未被很好体现。

同时页面最左边一列，把文献又进一步分为多个面，每个面即为可选择的筛选条件。但是知网对于文献类型的筛选栏也是相对固定的，都是分成了 9 大类，而且这些类型虽然能够较为全面的描述文献的各方面特征，充分揭示出各种主题因素之间的相互关系，能够实现多途径检索；但是对应多途径检索也一定程度上是复杂的，给用户检索会形成一定的负担，要求用户对于检索文献需要较为了解。

图 1-5　知网检索结果页面

2.2 主题组织法

主题组织法在当前时代下应用较为广泛。用户在进行文献检索的过程中，可以使用该文献的主题词作为检索的输入内容进行搜索，可以高效率得到检索结果；而当用户不能够清楚输入文献的主题词，以可以按照文献所具有的关键词进行检索，在进行进一步筛选也能够得到最后需要的文献。而选取的关键词一般需要出现在文献的标题、摘要当中效果较好，如果只是正文里才出现的关键词需要出现较高的频率才能够快速被用户找到，因此检索的效果要低于主题词的方式。

在使用前文构建的关键词进行文献检索的过程中，知网的检索结果是按照与关键词之间的相关度进行排序的。而突出的主题则大多数大学生就业质量、新时代大学生、新时代、就业质量是出现最多的。当然该系统也会给出一些与之相关的词语，比如高质量就业、就业价值观等词语。但是也会出现一些与之关系不强的词语，例如现状研究，现实审视、提升路径等等。

而在具体的检索结果之中，红色标注部分便是对于检索关键词的匹配，该本体部分主要采用主题组织法进行，根据用户输入的检索表达式进行检索，然后中国知网的内部系统直接将各个文章的标题中的与检索表达式对应的关键词进行标红输出，从而形成检索结果显示。此外，该部分排序主要还是依托于此前的排序方式的选择，一般默认为相关度降序的方式。其中红色标记越多的文献是与需要检索的内容匹配相关度是较高的，并且除了标题外，文献还有其他方面的描述；比如作者、来源，发表时间，被引、下载以及综合类的操作。可以看出除了对于文献篇名的匹配，还有对于来源的匹配。另外，采用不同的检索表达式得到的结果也是不尽相同的，大体而言文献来源的期刊是较为类似的，只是文献的篇名不一致。

图 1-6 表达式一结果展示

图 1-7 表达式二结果展示

2.3 时序法

通过文献发布的日期，知网将文献按照年份发展的顺序将文献进行排序，可以方便用户选择较新的研究成果与较早的研究成果，还可以将其进行对比分析。

2.4 体系-组配组织法

在知网检索的文献来源当中，并没有将各个文献来源进行严格分类，而是采用了体系-组配的方式进行展示。比如从期刊角度而言，原本期刊可以作为一个大类进行单独划分，但是知网却将学术期刊与特色期刊作为两个类进行表达，这样依据文献的功能属性进行划分的方式主要是组配的方式；而学术期刊和学术辑刊，则是按照学术文献背后所属的机构以及所蕴含的价值进行划分的两种形式，其他论文、会议、报纸、年鉴等分类则是形成了一个简单且全面的体系

2.5 检索效果评价

本次搜索累计得到结果为：

若是按照查询表达式一进行查找，最后能够得到 102 篇文献，但其中高度相关的文献却只能得到 51 篇，即准确率只有 50%，相对而言是比较低的。

而对于查询表达式二进行分析，最后能够得到 101 篇文献，但是高度相关的文献数量却只有 46 篇，准确率只有 45.5%，相较于第一次检索的结果还要低。

所以该检索系统的信息组织对于检索结果的准确率还应该进一步优化去提高，这样才能够更好满足用户需求。

3.信息组织描述

信息描述亦称信息资源描述，是指根据信息组织和检索的需要，对信息资源的主题内容、形式特征、物质形态等进行分析、选择、记录的活动。信息描述的结果是信息记录，也称元数据。而对于一篇期刊论文的元数据方案描述主要包含以下几个部分：

（1） 主题元素项

在期刊论文检索中，采用文献中的关键词作为检索入口；它是一种自由词。在选词时，应遵循准确性、科学性、使用频率高的原则，选取具有检索意义和名词性质的词汇。

（2） 作者元素项

该元素项应当包含作者的所属机构、职称、学历等属性的限定，有利于对论文作出相应的学术水平判断，方便联系交流。

（3） 描述元素项

论文往往受到一些机构所设立的基金项目的支持，一定程度上能从侧面反映论文的学术水平与质量,所以在描述元素中应该增加资助一项的限定修饰，包括资助基金名称、项目名称及项目编号。

（4） 资源类型元素项

对应于期刊论文则该元素项应明确为“期刊论文”。

（5） 标识符元素项

标识符是指在一定背景下可对资源做无二义性参照，使用一个字符串或数字识别资源，因此，有关期刊论文的参考信息，应包括充分的母体信息的期刊题名、ISSN号、年、卷、 期、页等信息，并不规定必须使用某一编码方案，可以文本的方式提供，在此扩充“文献引用”限定修饰。

（6） 相关资源元素项

一般为参考文献，著录参考文献能够表达对他人劳动成果的尊重，也是尊重知识产权的表现，同时也可以反映作者论文的真实科学依据和严谨的治学态度、提高文章质量，另外可以为读者提供文献线索，方便读者查阅检索。

（7） 论文类型元素项

为论文的属性，从其内容或形式进行分类，可根据正式的标识系统进行标识，通常可将论文类型分为综述性(述评性)、一般研究论文、研究简报等。

图 1-8 检索文献实例页面

从上述页面中能够清楚找到该文献的主题、作者、所属机构、关键词、摘要、专辑、专题、分类号等等元数据。

另外，在该页面的左侧部分清楚列出该文献的文章目录，让读者能够一目了然

整篇文献的大体框架以及背后所体现的逻辑思维。

而在文献主题的左上方是文献的所属期刊名称以及发行的时间，通过该期刊的名称能够一定程度上体现出该篇文献的学术价值的大小。

在文献主题的右上方则是对于文献相关操作的体现，主要包括引用、收藏、分析、打印、关注以及做笔记，便于用户认真阅读文献。

在页面的最下方则是进一步阅读文献的方式，可以选择手机阅读、也可以是网页阅读；还提供了文献的下载方式，可以使用 CAJ 格式下载，也能够导出为 PDF 下载。最下面提供了手机 APP 的二维码，方便手机用户同步阅读，也进一步展示了文献的下载量、页码、页数以及文献的大小等信息。

与此同时，知网还提供了核心文献推荐页面，将检索关键词的研究起点以及研究来源进一步罗列出来，经过节点文献进一步展示后续研究分支与研究去脉。

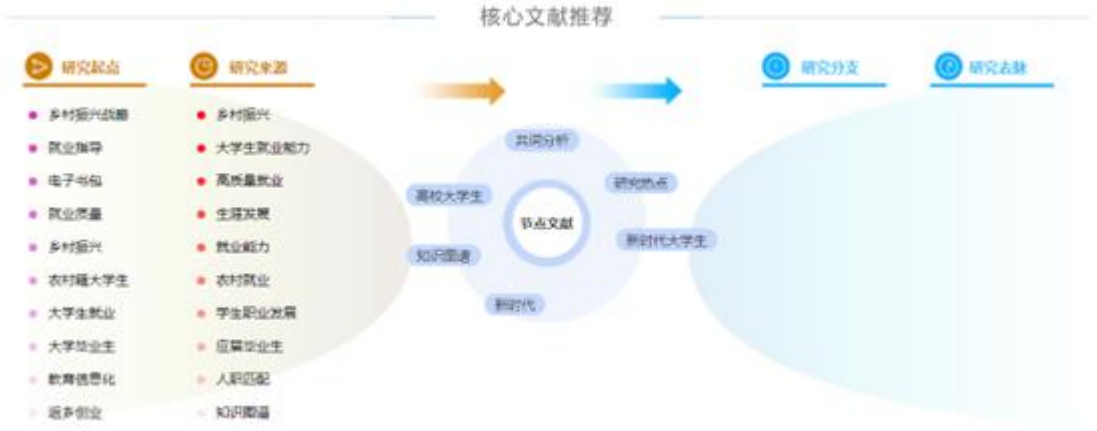

图 1-9 核心文献页面展示

另外，对于该篇文献的参考文献也进行了进一步分类。主要有参考文献、引证文献、共引文献以及同被引文献等等，并将这些文献进行具体展示。能够进一步为读者体现该文献的热度，提供参考价值。

图 1-10 文献参考页面展示

4.信息组织评价

4.1 用户体验角度

知网提供了多种检索方式，用户可以基于自己所擅长的方式自主选择，实用性还是比较好的，并且在网页设计的排版上来看也是较为条理清晰，脉络分明的。用户以自己的实际需求选择文献相关学科以及文献的主要内容形式，同样可以多方面同时组合进行筛选文献信息，简单实用，能够满足广大用户的需要。

但是在检索过程中，如果每一步都按照严格条件进行限制，那最后用户可能压根搜寻不到所要查找的文献，其次就像上述查询结果分析一样，最终得到文献的准确率相对而言是比较低的，一定程度上可能不能够满足广大用户的需要，并且对待一些比较不常见的知识领域，知网也不能够提供较多的文献。而知网提供的文献来源广泛，用户只能够依据自己的经验或者上网搜索文献的来源来评判文献的质量，并不能直观感受文献的水平，因此后续还需要进一步优化等级评价标准。

4.2 用户认知角度

知网的页面设计了众多的文献类型的选择，同时也提供了众多学习服务平台。用户可以根据自己所处的身份，所在的领域以及所要研究的事物类型进行链接的访问，非常便利。在学习课程知识的同时也能够根据用户的兴趣，选定特殊的领域，如汽车、建筑、电气、教育、研究生等等的平台进行访问学习；同时知网也提供了出版平台与评价，用户可以根据自己的实际需要进行访问学习，极大丰富了用户的认知，增强了用户的兴趣。此外，知网还提供了一些软件产品，以及相关热点资讯，用户可以及时掌握知网的动态消息。

但是，众多的用户在知网上的操作大多是进行文献的搜索以及相关论文期刊的下载，而忽略了众多学习平台的认知。对于其中特色的一些产品体验较少，也不会去花时间进行进一步体验。另外，对于知网的特殊产品在收费方面也是比较贵的，大多数用户是不会直接进行购买使用的。此外，知网对于这些平台的排版显得比较随意，有些页面的图片看起来好像是插入的小广告，让用户直接没有访问的欲望。

实验二

1．检索策略设计

研究问题	新时代高校大学生就业质量影响因素分析
研究目标	形成文献综述，大学生就业质量影响因素分析
时间范围	10 年
地域范围	全国
背景/事件	高校毕业生数量不断增加，就业形势严峻

2．提取检索词

university student,university graduate；

employment quality；

influence， impact

3．构建检索表达式

{(university student) or (university graduate)} and (employment quality) and {(influence) or (impact)}

4．选择检索信息源

在对 EI、ISI Web of Science、IEEE/IET Electronic Libraryz 等几个外文数据库进行初步检索尝试后，在检索界面、使用难易程度、检索结果数量等方面对几个数据库进行对比，发现 EI 检索界面相对亲和、容易上手、检索结果数量适中，最终选择 Engineering Village 作为检索信息源。

Engineering Village 涵盖了工程、应用科学相关的最为广泛的领域，内容来源包括学术文献、商业出版物、发明专利、会议论文和技术报告等等。其中的 Ei Compendex

数据库是美国工程索引 Ei 数据库，是全世界最早的工程文摘来源，世界三大检索工具之一。

5．构建查询表达式

(1) university student and employment quality and influence

(2) university graduate and employment quality and influence

(3) university student and employment quality and impact

(4) university graduate and employment quality and impact

6．信息组织方法分析

(1)分类组织法

EI 的检索方式中采用分类组织法，根据使用的目的和方法，将检索方式科学地划分为快速检索、专家检索、词典检索、作者检索、机构检索，直接给用户提供了多种检索文献的角度和途径，便捷好用。

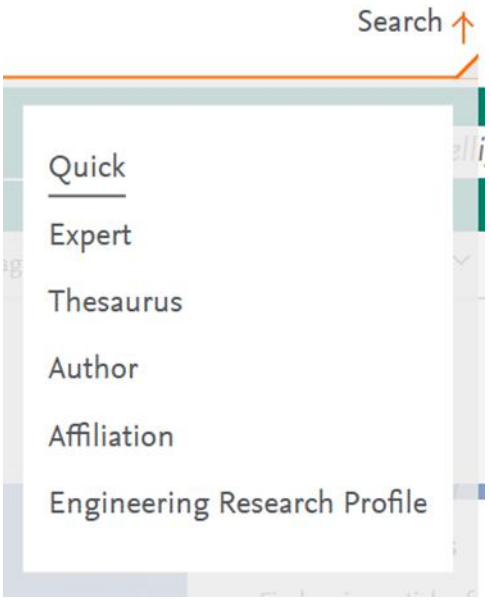

图 2-1：检索方式

EI 的快速检索的字段检索范围中采用分类组织法，除了所有这一范围以外，将字段检索范围科学地划分为主题/标题/摘要、摘要、作者、第一作者、作者从属、标题等多个范围，精准服务于用户进行快速检索，并使检索结果更加准确。

图 2-2：检索字段范围

EI 在检索结果限制的文献类型中采用分类组织法，根据文献的出版形式将文献分为多种类型，即除了第一个所有文献类型以外，剩余的所有文献类型，满足用户检索时对文献类型的特定需求。

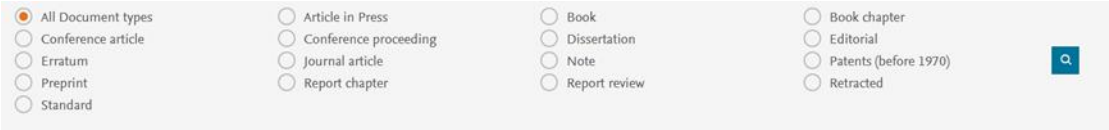

图 2-3：检索结果限定条件-文献类型

(2)主题组织法

EI 的检索界面顶层，采用主题组织法，使用标题词，将用户使用的关键功能置于顶层，主要包括检索方式、检索历史、警报、创建/登录和帮助这五个关键功能，其余功能在更多这一选项里面，方便用户快速选择不同的功能，增加检索系统的便捷性。

图 2-4：功能模块

EI 的快速检索中采用主题组织法，使用主题词，用户可选择一至多个检索字段，

限定相应的检索范围，选择 and、or、not 三种逻辑关系，从而实现快速检索。

图 2-5：快速检索

EI 的检索结果采用关键词法，将文献的标题、作者、作者从属、来源、文献类型、摘要等重要标识信息展现出来，同时对标题、作者进行标色加粗，对摘要中与检索字段匹配的相关信息加粗。检索结果直观地展现出表达文献主题的重要内容，利于用户对于文献主题的快速浏览，提高检索效率。

图 2-6：检索结果示例

(3)组配分类法

EI 的检索结果限定条件采用组配分类法，将文献科学划分为不交叉的多个组面，即文献的不同属性，包括数据库、日期、语言、文献类型、排序方式、论述类型等几个组面。用户可以按照自身需求，通过在不同的组面下选择限定条件，从而完成对检索结果进行限定。

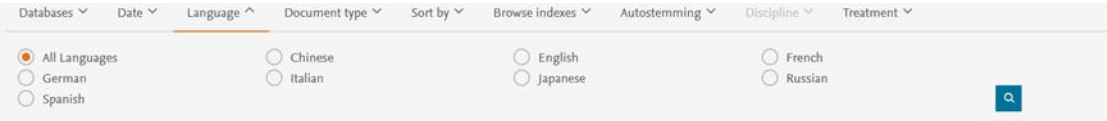

图 2-7：检索结果限定条件

EI 的精炼检索中采用组配分类法，将文献科学划分为不交叉的多个组面，即文献的不同属性，包括开放获取、文档类型、作者、作者从属、受控词汇、分类代码、国家/地区、语言等。用户可以按照自身需求，选择一至多个组面，并在相应的组面下选择限定结果，从而实现对检索结果的精炼检索。

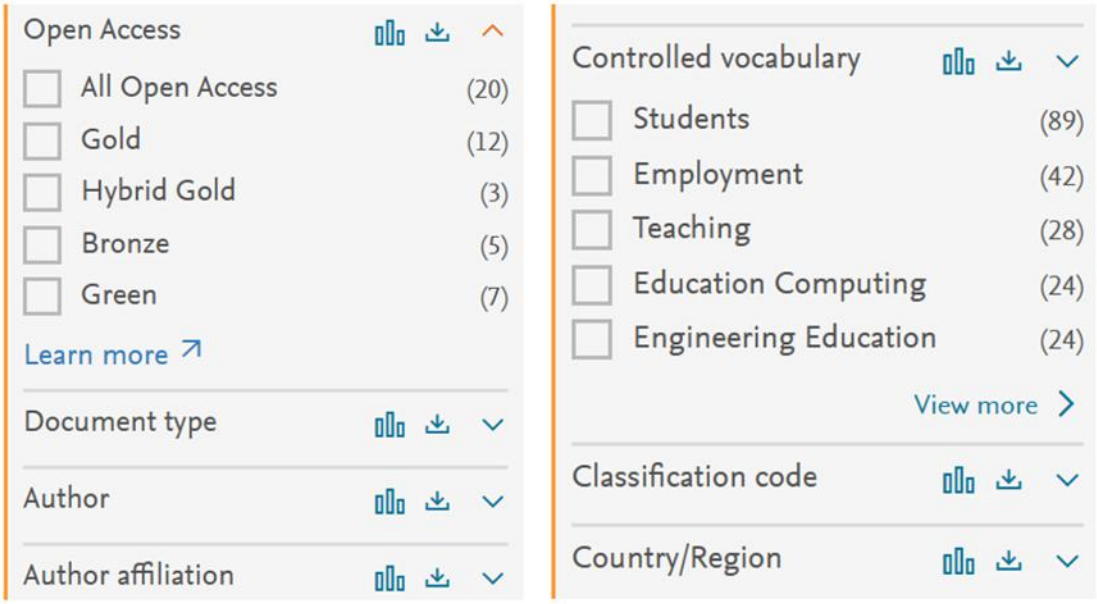

图 2-8：精炼检索部分

(4) 排序法

EI 的检索结果的排序方式采用排序法，系统默认按照相关度对检索结果进行排列，同时提供了时序法和对作者、来源、出版商的字顺排序法。用户可以根据自身需求在多种排序方式中选择一种，检索结果即按照用户选定的方式排序。

图 2-9：检索结果的排序方式

7．网页信息描述

信息描述是指根据信息组织和检索的需要，依据一定的规则和标准，对信息资源的主题内容、形式特征、物质形态等进行分析、选择、记录的活动。信息描述的结果，是一条有关该信息资源的书目数据记录，它由若干信息描述项组成。所以，信息描述实质是一个按照一定规则分析和选择数据的过程。

信息描述的结果是信息记录，也称元数据。元数据是关于数据的数据，是描述任何 Internet 数据和资源的数据, 是促进 Internet 信息资源的组织和发现的有力工具。

对于一篇会议论文的元数据方案描述主要包含以下几个部分：

(1)主题元素项

主题元素项是指资源内容的主题，通常是指描述会议论文内容的关键词、主题词或分类号。

(2)主要责任者元素项

主要责任者是指创建资源内容的主要个人或团体，即指会议论文著者，同时包括责任方式、机构等内容。责任方式是指责任者与会议论文之间的关系，主要包括撰写和翻译，机构是指论文责任者所在机构名称，通常用三级机构方式表示:机构名称、一级下属单位名称、二级下属单位名称。

(3)描述元素项

描述是指有关资源内容的描述。凡不能在其他专门的元素或元素修饰词项反映的有关会议论文内容的说明，包括论文的文摘、目次、资助等。资助是指论文研究所受资助的基金名称。一般指论文受到某个机构所设立的基金的支持，可以具体到项目名称及项目编号。

(4)资源类型元素项

资源类型是指有关资源内容的特征或类型，根据《专门数字对象描述元数据规范》子项目组规定，其值为会议论文。

(5)标识符元素项

标识符是指在一定背景下可对资源做无二义性参照，根据-一个正式的标识系统，使用一个字符串或数字识别资源。正式的标识系统包括 URI、DOI 等。

(6)相关资源元素项

相关资源包括参照和权限，参照是指所描述的资源参考、引用或者指向了另一资源，权限是指有关资源权限管理的声明，或者对服务机构提供该资源的参照。权限管理包括知识产权、版权和其它产权。如果该元素不存在，不能对权限管理作任何假设。

(7)论文类型元素项

论文类型是指根据论文的内容或形式对论文的分类,根据正式的标识系统进行标识，通常可将论文类型分为综述性(述评性)、一般研究论文、研究简报等。

下面以一篇会议论文文献的介绍界面为例进行网页信息描述。

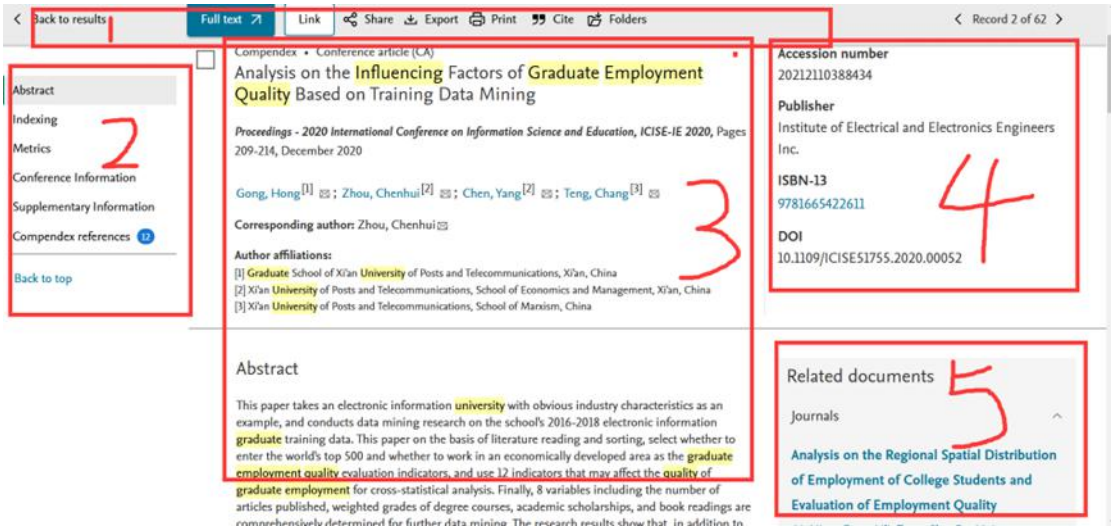

图 2-10：一篇会议论文的文献介绍界面

(1)图中的第一部分位于页面顶层，是对网页和文献的操作，包括返回上一级、全文链接、链接、分享、到处、打印、引用、下载这些操作，非常方便用户的使用。

(2)图中的第二部分是会议论文描述部分的目录，包括摘要、索引、指标、会议信息等，用户点击可实现页面快速滑动跳转。

(3)图中的第三部分是对会议论文的描述部分，包括主题、作者、摘要、索引、指标、会议信息、补充资料、参考文献，基本涵盖了所有元数据描述方案内容。全面展示出一篇会议论文的关键信息。

(4)图中第四部分是会议论文的收录信息，包括入藏号、出版商、国际标准书号和数字对象标识符。

(5)图中的第五部分是系统基于该篇文献推荐的相关文献，其中包括期刊论文、会议论文、新闻章节、书籍章节等类型，给用户的使用带来便利。

8．信息组织评价

8.1 检索结果评价

对于四个查询表达式的检索结果为：第一个检索到 90 篇文献，其中高度相关的有 16 篇，准确率为 16/90；第二个检索到 62 篇文献，其中高度相关的有 11 篇，准确率为 11/62；第三个检索到 118 篇文献，其中高度相关的有 20 篇；第四个检索到 82 篇文献，其中高度相关的有 26 篇。

从这四个检索结果可以看出，检索结果的准确率是相对较低的。分析其主要原因有两个方面，第一，由于英文单词具有多重含义，可用于多种情况的原因，导致查询表达式容易与许多不相关的内容产生一定的匹配度，从而大大降低检索的准确率。第二，本次检索采用的检索方式是 EI 的快速检索，采用多个检索字段匹配，其准确率本就较低，检索结果按照相关度排序，高度相关的文献几乎都居于较前的未知。

8.2 信息组织评价

(1)页面结构

EI 检索系统为功能模块专门设立一层页面结构，居于检索页面的上层，即功能模块位于顶层。用户在浏览页面下方部分的检索结果时可以快速切换功能，给用户带来便利，但同时，功能模块占据了部分空间，使得检索页面整体空间略小。另外，EI 检索界面的各项功能布局相对合理，使用方便。

EI 的页面颜色设置多为白色，部分灰色，给用户以较为亲和的感觉。同时 EI 的各项检索功能按键颜色均为深色，更加醒目，提高检索系统的便捷性；检索结果中每篇文献的重点内容的深色标识便于用户在粗略浏览文献的过程中快速获取重要信息，做出准确判断，增加检索系统的实用性。

(2)组织结构

EI 检索系统根据不同功能，设立相应的类目，包括检索结果限定、精炼检索、展开摘要等等。用户在使用这些功能的过程中，可以选择展开或收缩各功能下的类

目，对页面空间使用合理，给用户较好的体验。

⑶链接

对于检索结果中的文献，用户点击文献名称后可快速跳转至文献介绍界面。文献介绍界面的顶层的功能栏有返回上一级的选项，用户点击即可返回至先前检索结果已浏览到的部分，提高了链接的有效性和实用性，并具有不错的用户友好性。

在检索结果中，每篇文献都有两个颜色分明的全文链接在文献重要信息的最后一行，分布合理，提高了链接的实用性；两个链接中，一个是新建窗口，一个是直接跳转，给用户提供了多选择，提高了链接的实用性。

⑷导航

EI 的功能模块有一个“？”图标的帮助选项，其中有常见问题和使用教程视频选项，大大降低了用户使用该检索系统的上手难度，给用户带来较好体验，用户友好性较强。

⑸检索

EI 提供多种检索方式，快速检索、专家检索、词典检索、作者检索、机构检索，能够满足用户的多种检索需求。EI 在检索过程中提供自动控制词汇提示功能，提高了用户的检索准确性、高效性和便利性，为用户带来较好体验。EI 提供检索记录的共享功能，共享功能拓宽了检索的使用范围，用户可通过电子邮件或链接嵌入的方式将检索记录分享给目标对象，从而节省重复检索，促进信息流通，提高资源利用。

实验三

1.搜索引擎的选择

本文选择 Yahoo 作为研究对象。Yahoo 是全世界网络流量最大的网站，也是最早的门户网站。后来的大部分门户网站都是参照它的模式建立和经营，提供的网络广告形式也大都拷贝 Yahoo 的形式。Yahoo 有英、中 、日、韩、法、德、意、西班牙、丹麦等 12 种语言版本，各版本的内容互不相同。提供目录、网站及全文检索功能。目录分类比较合理，层次深，类目设置好，网站提要严格清楚，网站收录丰富，检索结果精确度较高。对于研究网络信息组织方法有较好的参考性。

2.信息组织分析

Yahoo 搜索引擎使用了多种信息组织方法，下面将一一详细介绍。

2.1 分类组织法

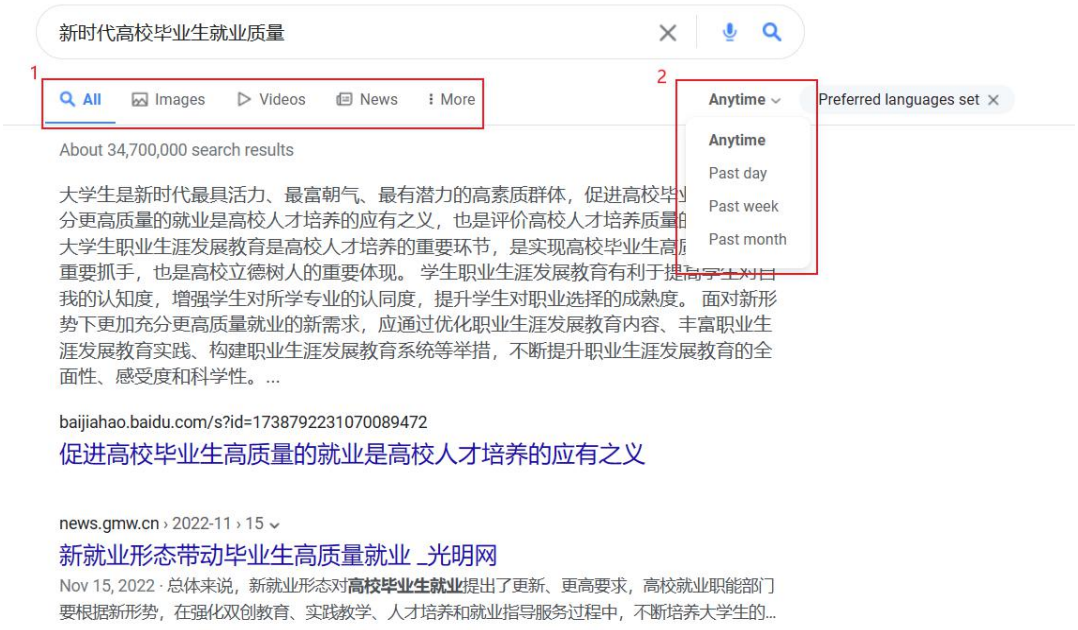

图 3-1：分类组织法

Yahoo 对信息进行了分类组织，将图片视频与文本数据等按类划分，具体分类

如图展示。可以看出，搜索引擎返回结果给出了两种分类方式：

在图片中 1 的位置，搜索结果被分为了图片、视频等形式，用户可以通过选择，查看文本数据以及多媒体数据。

在图片中 2 的位置，搜索结果可以按照时间进行细致划分，通过选择不同的时间段，检索结果将进一步的细化。

2.2 主题组织方法

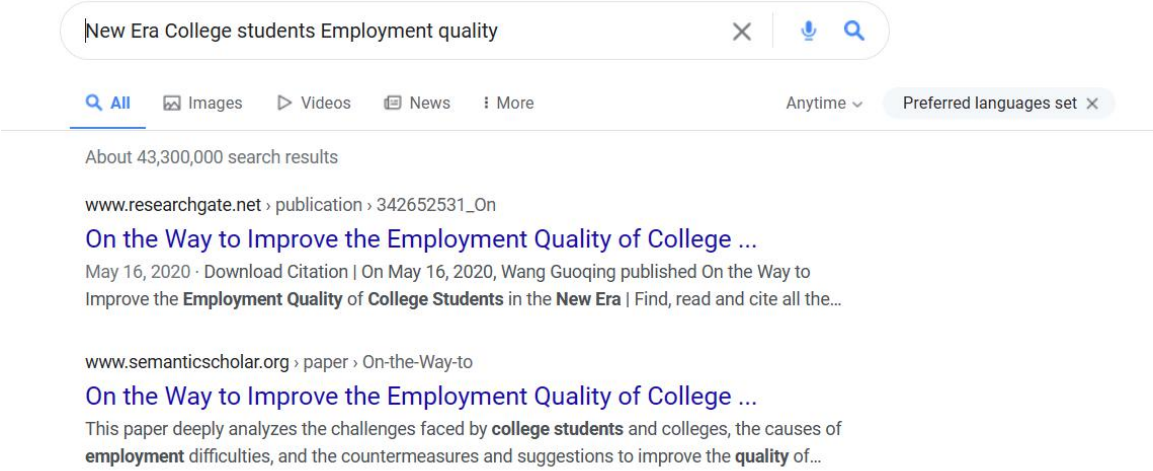

图 3-2：主题组织法

Yahoo 使用主题组织法对其数据进行组织，在搜索结果返回时，搜索系统按照检索表达式中的关键词，在数据库中进行匹配，最后返回结果，而匹配到的关键词则会在搜索结果中通过加粗加黑进行标注（如上图）。

2.3 排序法

Yahoo 搜索引擎通过返回结果中与检索词的相关度大小进行结果排序。

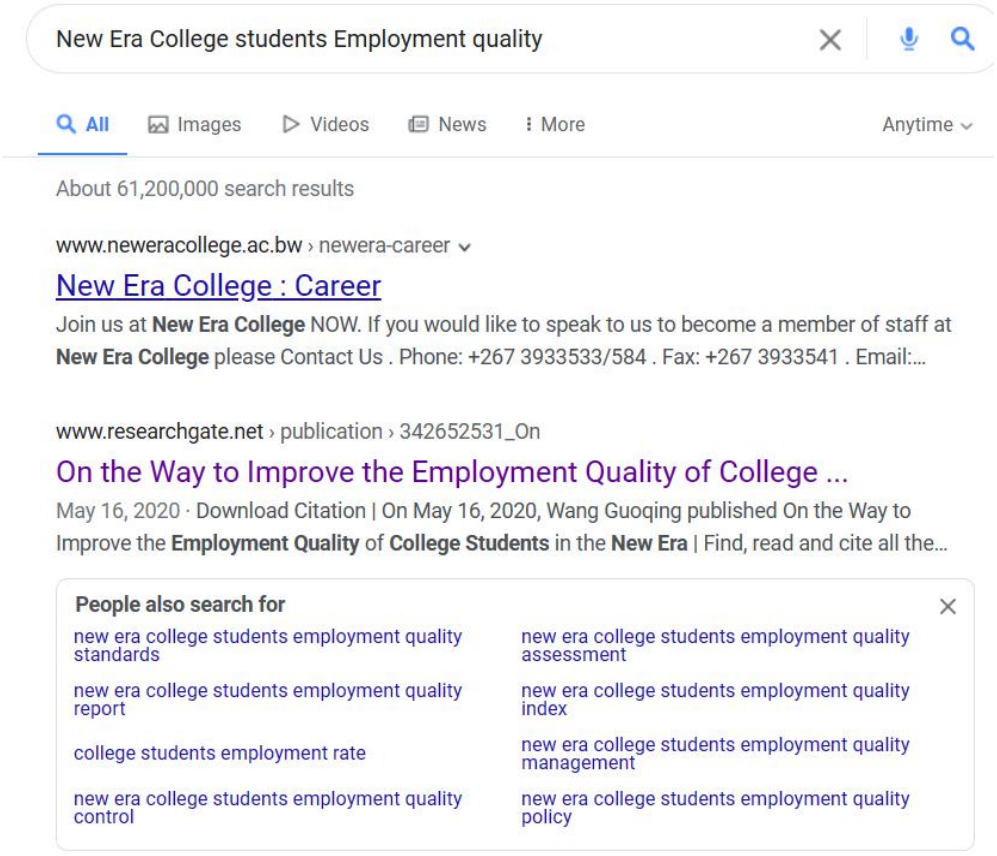

图 3-3：搜索结果排序

可以看出，排序靠前的网页的标题和简介部分和检索主题词的相关性更高。

2.4 关键词组织法

关键词组织法，除了体现在通过关键词进行检索外，还可以在图 3-3 的相关检索推荐中体现。

当点击网页的关键词包括 New Era College students Employment quality 等词后，Yahoo 会根据这些关键词出现在其他网页中的频率、重新进行组配，来组成新的检索语句，进行推荐检索。

2.4 信息描述

Yahoo 搜索引擎的信息资源包括大量的 PDF、DOC 等非网页资源，此类资源通过其原格式进行储存。

对于网页资源，Yahoo!提供 Catalogue 查询和通过数据库的 Web 查询两种检索方式，目录查询方式只检索 Yahoo 自己收集的网站资源，而 WEB 查询除此之外，还提供由第三方收集的网站资源，同时，支持布尔逻辑检索和进阶检索，可限定查询的资源类别、更新日期、显示数目等。

检索结果提供相关网站的概要描述和链接，以及该网站中符合条件的网页的标题、摘要及链接，并按分类类目及网站信息与关键字的相关程度进行排列。具体返回情况如图所示：

图 3-3：网页结果元数据

可以看出，网页搜索返回的元数据具体信息包括三部分，分别是图中的 1：网址导航，图中的 2：网页标题，以及图中 3：基于网页内容的概括性内容。图中 3 部分的内容，不是简单的内容概括，而是包括网页大部分内容的网页快照，可以在网页失效后，最大程度上的还原网页内容。

本部分重点参考了杨艳丽 [1]的硕士论文中对元数据的描述，但没有找到 Yahoo 网站使用 DC 与 RDF 的标准语言的应用实例。

正如上文所说，Yahoo 还提供了除音乐搜索外的很多搜索内容，在图片类型数据中，Yahoo 提供图片的来源、题目、图片大小、图片尺寸、图片格式、显示比例元数据。

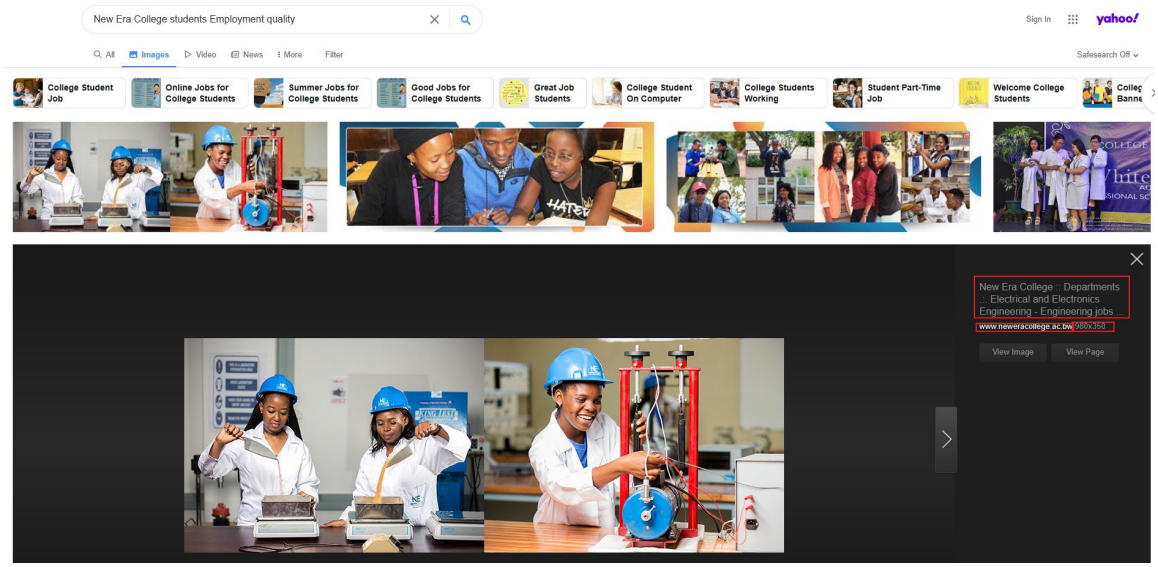

图 3-4：图片元数据展示

以及对于视频类型数据，包括题目、出处、时长、发布日期在内的元数据。

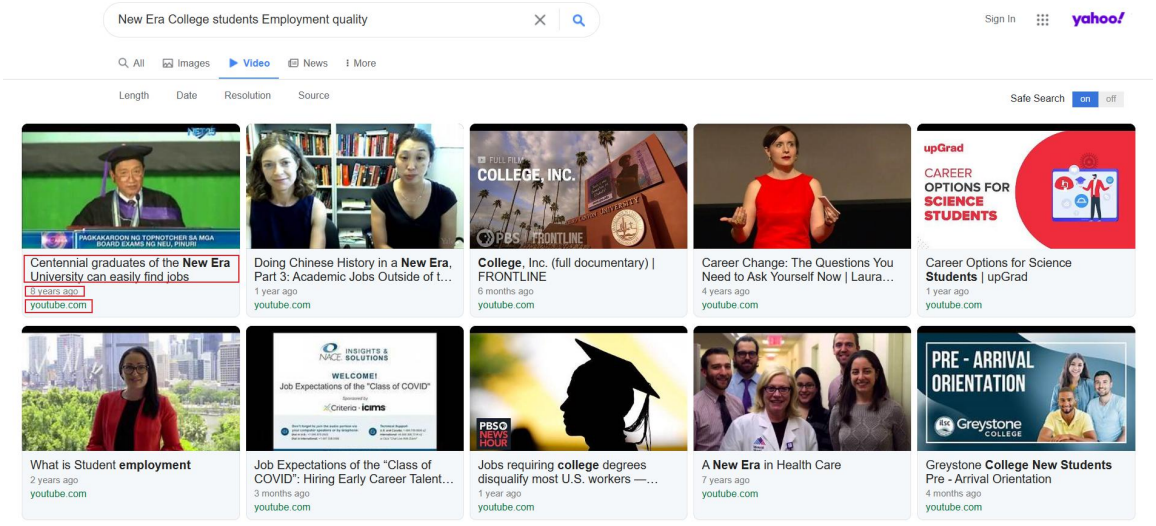

图 3-5：视频元数据展示

2.5 检索结果评价

在搜索引擎中，由于搜索结果返回数太多，难以全面计算，所以仅选择返回结果中的前 5 页进行准确率的评估（每页十条数据）。

选择中文、英文查询表达式的前两个的检索结果为：

第一个检索到 7 个相关网页，准确率为 7/50；

第二个检索到 8 个相关网页，准确率为 8/50；

第三个检索到 23 个相关网页，准确率为 23/50；

第四个检索到 17 个相关网页，准确率为 17/50。

从这四个检索结果可以看出，检索结果的准确率是相对较低的。而英文检索表达式的结果又明显好于中文表达式，这可以理解为 Yahoo 收录的英文网页要远远超过中文网页，才导致了类似意义的关键词返回的中英文差异如此之大。

3.信息组织评价

3.1 资源收录状况

（1）收录的“全”

在数据的收录方面，本文将从百度和 Yahoo 的搜索结果对比来说明。

下面，以“新时代高校毕业生就业质量”作为关键词进行检索，这两个搜索引擎得出的网页结果差距很大。上文提到的，以该搜索词进行的检索，Yahoo 搜索引擎得到了 34,700,000 条返回结果。而百度搜索中，返回结果仅有 13,200,000 条

图 3-6：百度引擎返回的搜索结果

可以明显看出，Yahoo 收录的相关数据要远远多余百度搜索引擎，虽然百度并没有给出权威的网站收录量，但在相关资料中显示，Yahoo 的强大数据库得到了更多人的认可，收录的网站数也公认的多于其他搜索网站。

在图片搜索方面，Yahoo 也以绝对的数量优势压倒性的打败了百度搜索：

	图片数量	图片过滤	站内搜索	高级搜索	分类目录	新闻图片搜索	图片格式
Google	10 亿多	有	有	有	无	无	JPG/ GIF/ PNG
Yahoo	30 亿	有	无	无	有，分类详细	无	无
百度	近亿	无	有	有	有，分类极详细	有	JPG/ GIF/ PNG/ BMP

图 3-7：搜索引擎的图片搜索对比

上图展示了在 2009 年左右的相关搜索引擎的收录情况，虽然百度搜索在用户体验方面有更多的重视，但在各项数据收录上明显不足，Yahoo 在网站收录方面更全。

此外，Yahoo 作为顶级的搜索引擎，支持多语种搜索，对于其他语言的支持度也很高，使用英文检索表达式，得到的检索结果如上文所示，有 43,300,000 条。同时，对于不同类型的数据也收纳广泛，如 doc、pdf 格式的文档，Yahoo 收录也不在少数。

（2）收录的“精”

下面是网上给出的网页被 Yahoo 收录的最低标准：

（1）如果是商业网站，网站必须具有正式的商业名字，并在网站显著位置显示。

（2）网站必须定位明确。

（3）网站没有被 Yahoo 目录收录。

（4）保证所递交网站，其内容在 Yahoo 目录里是“惟一”内容。比如，您已经向 Yahoo 递交了一个网站 A，您又申请了一个域名，并又建立了一个网站 B，网站 A 与网站 B 内容相同或“换汤不换药”（虽然语言上有些改动，但实质上还是一个内容），这时您就不能再向 Yahoo 递交网站 B。

（5）如果网站是有地域特征的网站，必须有详细的地址。

（6）没有‘正在建设网页’。

（7）您的网站链接全部有效，并指向相关内容。

（8）网站必须是英文网页，或者有英文版。

（9）您的网站兼容多种浏览器，比如，不是纯 Java 网站。

（10）您的网站必须 24 小时与互联网相连。

可以看出，网站收录时，Yahoo 提出了较严格的要求，这种要求无疑提高了 Yahoo 搜索提供网站的精度，同时，如果是商业网站，Yahoo 有更高的要求：

首先，需要向 Yahoo 支付 299 美元(成人内容或服务网站需支付美元$600，并且在 Business and Economy/Shopping and Services/Sex 下的适当目录申请)，本部分资金，不管收录与否，Yahoo 官方都不会退回，这无疑提高了商业网站的入驻门槛。此外，还有以下要求：

(1) 但即使是您支付了美元$299，也不保证您的网站一定被 Yahoo 收录；

(2) 即使网站被 Yahoo 收录，也不保证是您递交网站时所选择的目录；Yahoo 工作人员 有权更改目录；

（3） 即使网站被 Yahoo 收录，也不保证是你递交网站时所填写的注释，即网站说明。

（2）、（3）项要求，虽然对网站有所修改，但提高了对网站内容的管理，对网站的精度有进一步的提高。

（3）收录的“快”

在数据收录方面，中文网站方面，百度搜索具有本土优势的天然优势。综合评价来说，Yahoo 对网站的收录速度是仅次于 Google 搜索的，并且优于百度的收录速度。但在内容更新方面，虽然 Yahoo 每天都会对网页进行更新，但对某一网站的更新，可能会存在不到 1 周的延迟。

同时，Yahoo 收录的“快”也可以在（1）中的实例中体现。对于“2022 世界杯 吉祥物”这个关键词，是最新的热门话题，可以看到，Yahoo 搜索引擎已经收录了巨大乃至庞大的数据，可以体现 Yahoo 收录的“快”。

3.2 数据的质量评价

上文已经从收录的精度方面说明了 Yahoo 数据的精度，下面将从单个搜索和百度搜索的对比中具体说明数据的质量。

首先是百度：

图 3-8：以 R 为关键词的搜索结果（百度）

其次是 Yahoo：

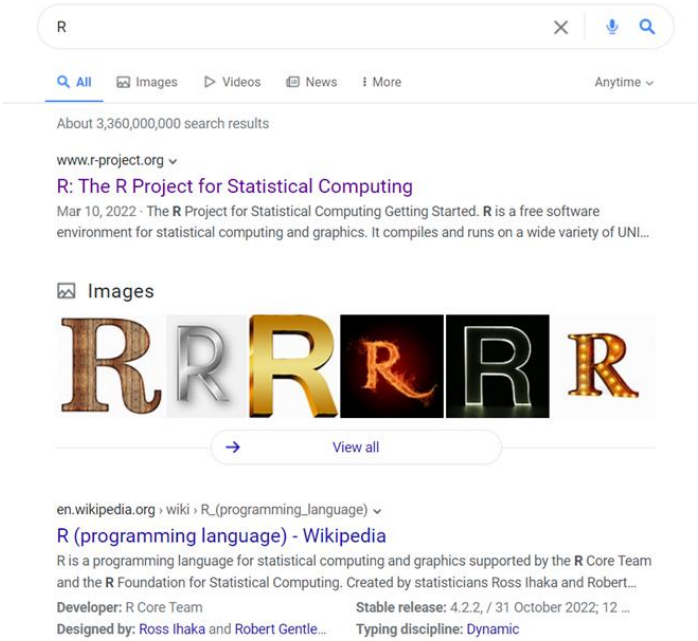

图 3-9：以 R 为关键词的搜索结果（Yahoo）

从上述搜索结果比较中可以看出，百度搜索提供的返回结果更多的是对 R 本身的解释，同时将自己的相关网页置顶，并在第四个检索结果中就出现了广告。而 Yahoo 则更重视返回结果的质量，将 R 语言的首页放在了首位，结果更具学术性、专业性，同时也没有广告的出现，检索的数据质量不言而喻。

3.3 检索的功能和效率评价

3.3.1 检索功能及特点

Yahoo 提供了十分强大的检索功能，下面将选择其中较为特色的几个方面进行讲解：

（1） 网页快照

收藏快照就是在用户收藏网页时，将网页中当时的纯文本内容抓取并保存下来，这样当这个网页被删除或链接失效时，用户可以使用收藏快照来查看这个网页的主要内容。注意：仅为纯文本内容，对于页面中包含的图片、Flash、音乐等非文本信息，快照页面还是直接从原来的地址调用。所以如果这些地址失效，那么快照上的图片等非文本内容将无法显示。

（2） 旅游搜索

以旅游景点作为检索词，会提供了景点介绍以及旅游攻略，可以点击提供的相关问题看到更多相关信息。此功能对国外景点有更丰富的介绍，包括相关旅游景点介绍、旅游攻略介绍等。

（3） 人物搜索

人物搜索是以网页搜索为计算基础，基于传统搜索的信息挖掘技术，将人和人之间的关系抽取出来，针对知名人士给出围绕知名人物的丰富信息。包括人物简介、人物关系等娱乐化内容。并可以根据指定关系查看详情。

（4） 股票搜索

在搜索词为公司企业、股票代码时，Yahoo 搜索会自动将股票相关信息置顶，直观的展示股情，很大程度提高了查询股票信息的效率。对于外股，还有详细的走向图等直观展示。

3.3.2 检索效率及评价

经过一段时间的实验，通过多个关键词的检索，发现 Yahoo 的前 10 页的查准率十分高，几乎没有明显的广告或不相关内容。但同时也发现，由于每个关键词内容，Yahoo 都给出了庞大的返回结果（即使是很离谱的搜索关键词，也会有大量返回），因此本部分不方便使用所学的正确率、召回率等评价方法对搜索引擎的搜索结果进行评价（好像机器学习可以进行 Precision 等的计算但很繁琐），同时，搜索引擎的评价涉及到很多方面，仅使用上述方法评价并不完整，下面三种评价方式均来自文献阅读与整理。

⑴ 相对查全率

查全率 (R) 是指检出的相关文献数量与系统中全部的相关文献数量的比率。由于无法精确确定因特网所有相关文献的数量，于彩云 [2]采用相对查全率 R (Ai) 来对 Yahoo 的检索效果做一评价。

具体做法为(aij 表示第 i 个搜索引擎对第 j 个搜索词返回的内容)使用 Abest ($a1^*$, $a2^*$, …, an^*) 记录对不同检索课题的不同检索词，查询时返回记录数均为最多的搜索引擎，用 Aworst ($b1^*$, $b2^*$, …, bn^*) 表示查询时返回记录数均为最少的搜索引擎。$ai=\Sigma(aj^*-aij)$ (i=1, 2, …, m) 刻划了搜索相擎 Ai 在查全方面与 Abest 接近的程度，ai 越小搜索引擎查全性能相对越好。而 $bi=\Sigma(aij-bj^*)$ (i=1, 2, …, m) 反映了搜索引擎 Ai 在查全方面优于 Abest 的程度，为此定义搜索引擎 Ai 的相对查全率如下:

$$R(Ai)=bi/(ai+bi) \quad (i=1,2,\cdots,m)$$

通过此方法，作者得出以下数据:

搜索引擎检索词	Data mining	Search engine	metadata	James Joyce
Yahoo (A1)	40 000 000	438 000 000	20 400 000	19 500 000
Altavista (A2)	37 000 000	435 000 000	28 400 000	17 800 000
Google (A3)	5 909 000	104 550 000	4 006 000	2 872 000
Lycos (A4)	1 633 171	120 785 756	1 617 143	1 113 664

图 3-10：相对查全率

并得到 Yahoo 搜索引擎的查全率为 0.9808。

⑵ 查准率

在查准率分析中，根据韩圣龙等[3]的研究(主要针对 Yahoo 中文，且时间较早)，可以看出实验一的 P(20)曲线明显高过其它四个实验的 P(20)曲线,这说明中文 Yahoo 检索出的能够满足检索提问式的记录当中有用的不多。在进行相关性检验时

已经注意到,检索结果当中有很多是产品介绍、科研机构介绍、科研人员介绍等毫无学术科研价值的信息。这也反映了当前网上中文信息资源的一个不足之处。

5 个实验的 P(20)平均值分别是 0.242、0.064、0.019、0.069、0.023。两组对比实验中,实验四和实验五的 P(20)平均值比实验二和实验三的 P(20)平均值略大,说明中文 Yahoo 的前 20 个命中记录中也有重复记录。实验二的 P(20)平均值是 0.064,说明中文 Yahoo 平均在前 20 个命中记录中能提供一到两个有用记录。

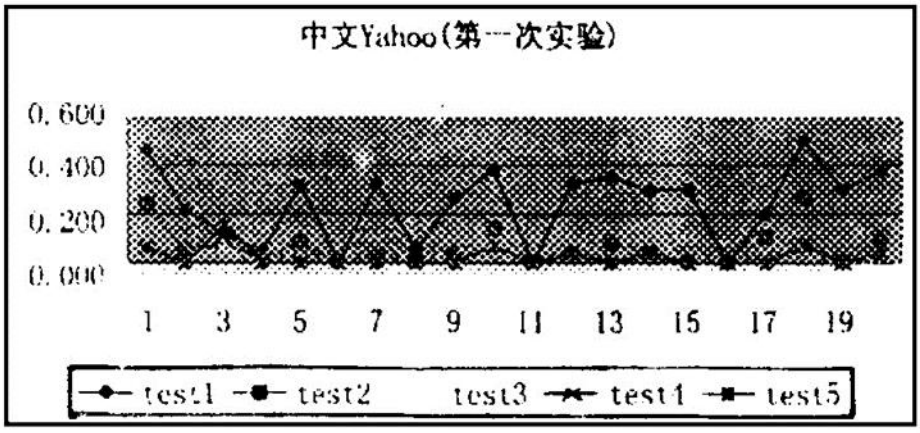

图 3-11：查准率

而在于彩云 [2]的研究中，“data mining in the digital library”这个命题在 Yahoo 搜索引擎中的相对查准率为 35%，可以看出，在这些研究进行时，Yahoo 的查准率并不高。

但在目前的用户体验来看，Yahoo 的查准率要优于百度搜索。

⑶ 响应时间

根据付天香 [4]从明星、壁纸、卡通、风景等角度来选择 10 个含义明确的检索词，进行的检索的实验，有如下结果：

	张娜拉	中国结	加菲猫	百变小胖	梅花	越狱	宝马	橘子	沙漠	日出
Google	0.03	0.02	0.02	0.06	0.03	0.04	0.06	0.03	0.04	0.02
Yahoo	0.002	0.002	0.005	0.002	0.005	0.002	0.002	0.002	0.002	0.002
百度	0.007	0.006	0.005	0.008	0.007	0.004	0.009	0.008	0.005	0.005

图 3-11：响应时间

有以下结论：

	图片数量	响应速度	检准率	结果去重	检索结果排序标准	另类检索方式	检索结果显示格式
Google	多	快	一般	有	链接点击频度排序	高级检索	缩略图、文件名、图片类型、像素、文件大小、图片的 URL
Yahoo	很多	极快	很高	无	文件名中出现检索词靠前	分类目录检索	缩略图、文件名、图片类型、像素、文件大小、图片的 URL，加入相册
百度	极多	很快	较高	有	文件名中出现检索词靠前	高级检索分类目录检索	缩略图、文件名、图片类型、像素，不包括图片的 URL

图 3-12：综合结论

3.4 检索评价其他方面

除了通过上述的权威方法对检索进行评估外，本文还从以下几个方面对检索进行了更加细致的评价。

(1) 检索入口

“虚拟的信息集合”是 Yahoo 的一大优点，体现在其拥有的概念模式和引用次序(即分面排列次序)的灵活性上。在传统的图书馆中，一本书只能放在书架的某一固定位置上。但在数字化的世界里，电子信息资源却不用再限制在唯一的物理位置上。我们可以将某一信息源分到类目结构的不同位置上。通过将分面分析方法应用到网络信息资源的组织中，Yahoo 能够为某一信息源在其巨大的分类等级结构中提供不同的路径分支入口，这样就使其能够从不同的路径，为检索相同内容的不同用户提供服务，从而完成查询。

例如，如果我想搜索“北京清华大学”的相关信息，Yahoo 搜索可以从多种入口进行搜索。

i. 从“地区”出发，先在地区内选择北京，再在北京的范围内搜索清华大学的相关信息。

ii. 从“教育”出发，先在教育类中选择大学/学校，再在学校中进一步缩小范围，搜索北京大学。

这种搜索方式也是 Yahoo 分类主题法应用的特有方式，大大提高了检索效率。

⑵ 检索组配

对于搜索引擎的组配搜索情况考察，是评价搜索引擎中的很重要的一环，词语组配如果没有搞清楚，那么对检索的精度将造成重大损失。

在使用多数有组配歧义的词语进行测试后，发现 Yahoo 搜索在此方面做的很好，不存在搜索结果不符合预期的情况。同时，Yahoo 等众多搜索引擎也提供了“”作为减少此类问题出现的措施，很好地解决了此类问题。

本文想着重介绍的是在检索过程中的模糊词语问题，如同汉语的一次多义、一语双关一样，英语也存在此类问题，如：The passerby helped dog bite victim 谷歌翻译为：路人帮助狗咬伤者，但其实并非此意，而是 bite 作为定语来修饰 victim。而在相关搜索结果中，这种存在模糊含义的句子或词，都以正确的含义进入了搜索引擎并返回了想要的结果。

⑶ 链接可用性

随着互联网的逐渐发展，越来越多的网页资源涌现而出，而有一些“年久失修”的钉子户，仍占有着推荐资源。这种网页在百度搜索中非常常见，在搜索电影资源、计算机编程资源时，经常会有域名丢失、个人博客无法访问、网页 404 的无效链接

出现，本文随机选取了在百度搜索中返回结果中有无效链接的 10 个关键词（中英文各五个）进行测试。在 Yahoo 搜索的返回结果中，仅有三个关键词存在无效链接，并且都出现在中文关键词的中文网站中。对比结果非常明显------Yahoo 搜索具有很高的返回结果有效性。

⑷ 用户界面

最后，本文对检索的用户界面进行了评价。

选取的搜索关键词为 Yahoo 搜索的热门词推荐。具体返回结果如下：

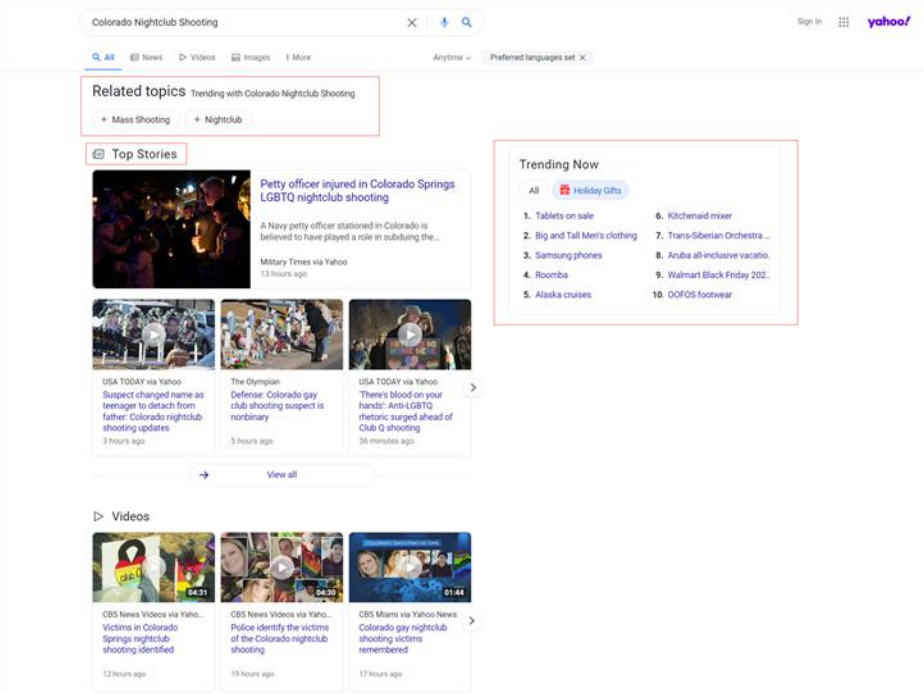

图 3-12：用户界面

可以看出，Yahoo 提供了较为简介明了的用户界面，广告很少，同时，提供了相关搜索推荐。在搜索返回结果的首页，还有图片内容、视频内容的相关推荐，方便用户使用。此外，返回结果的链接和网页没有大量的文字描述，言简意赅。尽管返回结果很多，但内容上没有重复内容，避免了用户重复打开网页不同但内容相似的网站。

4.实验三主要参考文献

[1] 杨艳丽. 元数据与网络信息资源的管理[D].太原理工大学,2003.

[2] 于彩云.搜索引擎 Yahoo 的性能评价及评价指标的选择[J].现代情报,2007,(02):185-187.

[3] 韩圣龙,赖茂生.网络信息检索工具评价实验(Ⅱ)——中、英文搜索引擎检索评价实验[J].情报科学,2001,(04):430-434.

[4] 付天香.Google、Yahoo 和百度的图像搜索比较[J].图书馆学刊,2009,31(02):103-106.

实验四

1.选取网站

本实验选择的网站是苏宁易购官网：https://www.suning.com

图4-1 苏宁易购首页

2.信息组织分析

2.1 网站首页的信息组织方法

苏宁易购首页是典型的一般网页，在网页的左边是对所有商品的分类，按照商品的本质类型分为了十五个大类，以及四十五个小类，从图中的类目来看，这些类目之间没有交叉，最细分的类之间也没有从属关系，属于体系分类法。

图4-2 苏宁首页的分类方法

将鼠标放到空调一行上，可以看到右边出现了对类目更加细致的划分，如将空调按照摆放形式分为挂机和柜机，按照空调的特殊功能划分了变频、新风空调、移动空调，按照制冷量的大小分为 1 匹、1.5 匹、2 匹、3 匹，按照品牌划分为美的、格力、海尔、TCL、海信、科龙、华凌等。这里可以明显看出来根据空调不同属性划分，这些属性之间可能交叉，即可能有冰箱同时满足多个标签，所以这里用了举典型的分面组配分类法。而排序方法较为混乱。

还有一些像食品这一栏下延续了体系分类法，比如休闲食品分为饼干、零食、面包、巧克力、坚果、蜜饯、糕点、肉松肉铺、膨化食品、糖果、薯片、礼盒，其分类如图 4-3，这些小类相当于第三级类目。

图4-3 食品栏下的分类

经验证，十五个大类中厨房小电、家具、食品、母婴、美妆、服装、滋补保健、清洗维修这八行使用了体系分类法的延续，其余的均使用了分面组配分类法。

2.2 苏宁易购的索引规则

苏宁的索引规则。也就是商品怎样才能被搜到，商品若想被正常召回，有以下五个因素：

1.商品上架状态，2.商品基础信息完善，3.商品有销售范围，4.商品有价格数据，5.商品有库存数量

上架的商品基础信息完善（包括标题、核心参数、卖点、亮点、主图等等），有明确的售卖范围，有明确价格数据，商品有库存数量。但是如果商家在商品标题里放好多其他品牌的词，会遭受在途的流量蒸发甚至是下架打击。从图 4.4 可以看到，苏宁所有商品都会明确指明品牌、品类、数量和质量等信息，不会出现多个品牌名称集于一个商品的情况。

图4-4 商品会明确指明品牌、品类、数量和质量等信息

商品能被检索到还有一个重要因素，即相关性，共有三点：

（1）文本相关性

商品的文本描述信息（包括：商品名称、标题、类目名称、品牌、重点属性，图书类商品还有作者、出版社）和搜索关键词的匹配程度。

（2）类目相关性

系统计算商品所在类目与关键词的相关程度，进行二次匹配，提高相关分类商品的权重。

（3）品牌相关性

计算商品的品牌与关键词的相关程度，进行二次匹配，提高相关品牌商品的权重。图 4-5 是搜索“苹果手机”的结果，能看到排在前面的都是苹果 14plus 或者苹果 14 既满足了文本相关性、品牌相关性和类目相关性，又符合用户最新的搜索购买趋势。

图4-5 搜索“苹果手机”得到的结果

由以上检索结果可以看出苏宁易购的检索信息是可以随时间更新的，经验证，具体的搜索信息更新方式有以下三种：

（1）实时更新

商品价格信息、商品主图、商品标题、商品卖点。

（2）增量更新

商品上下架状态、商品库存状态、促销信息。

（3）全量更新

凌晨（2 点以后）一般在早上更新结束，包含排序位置，评价信息，商品属性，结果页高筛等。

2.3 苏宁易购的排序规则

符合索引规则的商品能够保证召回，但召回的商品如何排序也有规则。图 4-6 是搜索“鸡尾酒”得到的结果，可以看到，苏宁易购平台的商品一般按照四个维护排序分别是综合、销量、评价、价格，默认按照综合纬度排序。

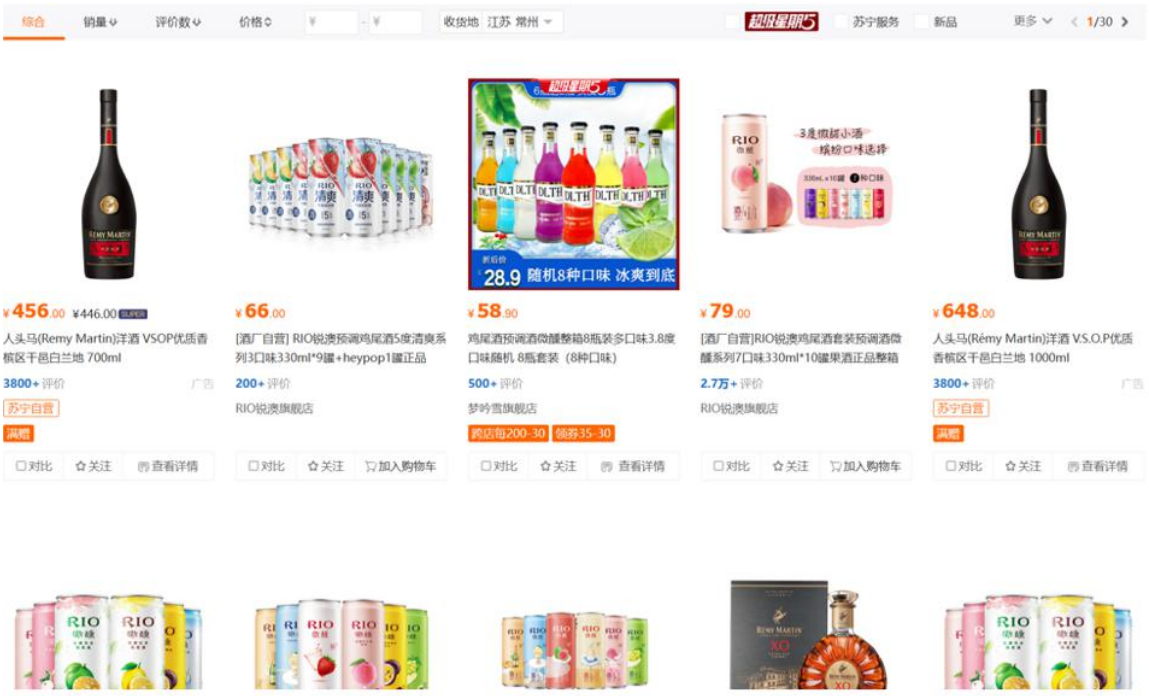

图4-6 搜索“鸡尾酒”得到的结果

单维度排序：按照销量、评论、价格单一维度排序方式。与前文中提到的关键词与商品类目的相关性有很大的关系。使用价格/销量/评论等非默认排序时，系统将相关性较差的商品类目过滤，不予以展示。比如上述结果中可以看到几个不是鸡尾酒的商品，但当切换成按销量排序时，检索结果就全部是鸡尾酒了。

对于综合排序的方式下商品的排序规则，苏宁易购主要依赖于以下 6 种因子：

（1）商品质量计算影响因子

销量：滚动周期内的（3 天、7 天、15 天、30 天）商品销量，虚假交易销量不计算。

订单量：包含该商品的订单数量。如某订单商品 A 销量 100，但该商品的订单量仍记为 1。

销售额：商品的销量数量*单价。

评价数：商品近 30 天内累计的用户评价数量。

好评率：商品近 30 天内用户评价数中好评数所占的比例。

收藏数：商品累计被加入收藏夹的次数。

缺货率：商品近 30 天内曝光次数中，无货的 PV 所占的比例。

加购数：商品近 30 天内被加入购物车的次数

(2) 用户反馈分计算影响因子

包括用户点击量、点击转化率、购买转化率，用户反馈表示用户搜索关键词后点击或购买商品的行为，反映了用户对搜索结果的满意度，同时反映了对商品的满意度，对搜索排序得分计算有重要影响。

(3) 店铺质量分计算影响因子

商户的服务质量对苏宁易购整体的体验有着举足轻重的影响。苏宁平台店铺综合得分统计包括以下五个指标分类：客服及售后、营销推广、物流时效、运营提升、增值服务。由开放平台计算，搜索系统获取最终店铺得分。

(4) 商品评价计算影响因子

有评价数、图片评价数、视频评价数、好评率、评价率、差评数、差评回复数、差评回复时效等指标。

(5) 个性化计算影响因子

有用户端标签和商品端标签两种。

(6) 反作弊手段

苏宁易购对于商家作弊以影响排序的行为拥有一定的反作弊手段，搜索作弊系统会对恶意作弊行为进行处罚，降低作弊商品权重，甚至于影响店铺 DSR。一般常见的作弊行为有以下几种：

低价刷单：原价 1000 元的商品，改成 10 元大量下单

用券刷单：店铺给自己发大额度促销券，支付少量现金购买商品

恶意点击：同一时间段，同一 IP，大量的点击行为

刷销量：系统根据模型识别出作弊销量，对该部分销量进行剔除，同时对商品进行处罚。

苏宁对这些作弊的行为识别的维度包括：IP 地址异常，页面停留时间异常，销售数量异常，成交价格异常，订单备注信息异常，收货时间异常，会员订单数异常，

收地地址异常，涉嫌刷云钻异常和邮费异常。

还有的商家可能会进行文本信息方面的作弊，比如商品参数作假、评论界面刷好评等，苏宁对这种作弊主要有两种防范方式：

关键词堆砌：检查某商品标题，如果标题内所写品牌名称和该商品参数内品牌不一致时，视为命中；如果标题内所写品牌词语数量超过 1 个时，且所写品牌和该商品无关，视为命中；

标属不一致：检查商品标题，标题内所写属性描述词语如果和商品参数内所写属性明显相悖时，视为命中。例如，标题内写男性衬衫，但参数属性中写为女性用品，则视为命中。

2.4 网站从页的信息分类方法

点击空调进入网站从页，其分类样式如图 4-7：

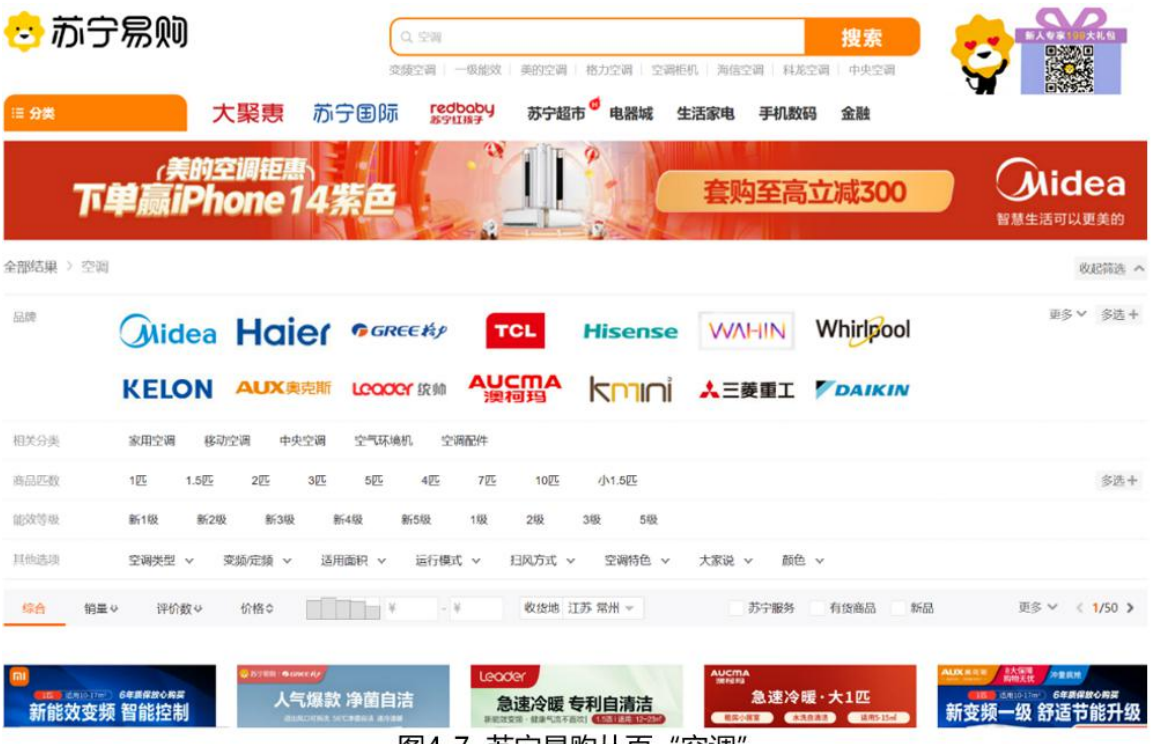

图4-7 苏宁易购从页“空调”

可以看到细分类用了标准的分面组配式分类法，一共分了十二个亚面，分别是

品牌、相关分类、商品匹数、能效等级、空调类型、变频/定频、适用面积、运行模式、扫风方式、空调特色、大家说、颜色，其中匹数、能效等级是按从小到大的方法排序，品牌有单独列出来按字母顺序筛选，但没有按照某种规则排序。其他的类也均没有固定的排序方法。

图4-8 空调品牌的排列

这个网页在每个亚面下进行选择，下方会出现满足结果的所有商品。

我选择了海尔品牌、2 匹、变频的家用空调，图 4-9 是呈现的结果。

图4-9 选择“海尔”品牌、“2匹”、“变频”后的结果

还有一些细分商品延续了体系分类法，比如家具板块：家具馆 (suning.com)，其页面如图 4-10 所示：

图4-10 细分商品“家具”的分类

网站的左边给出的是所有家具的分类，首先是按用途将所有家具分为了一级类目“客厅家具”、“卧室家具”、“餐厅书房”、“成套家具”、“儿童家具”、“定制家具”，每一类下依旧按照用途分为了更细的类，比如客厅家具分为了沙发、电视柜、茶几、客厅家具套装这四个二级类目，每个二级类目下还有三级类目，比如客厅组合分为茶几电视柜、沙发组合、沙发电视柜、沙发茶几。所以这部分的分类是典型的体系分类法。

总之在从业中只有厨房小电、家具、母婴三个用了体系分类法，食品、服饰、运动、箱包、医药五个用了体系-组配分类法，以体系分类为主，组配为辅，而且组配都是以品牌作为第二个亚面，剩下的是组配法，均有超过三个亚面。

3.个性服务组织方法

3.1 主页上的个性推荐服务

主页上的个性推荐服务如下：

图4-11 苏宁首页的推荐服务

最上方有几张与活动有关的站点，排版较为清晰，但是没有逻辑性可言，也没有按照某种规则排序。

网站中间的推荐服务排版如图 4-12、图 4-13 和图 4-14，可以看到从上到下依次是限时秒杀、领券中心、频道广场、猜你喜欢，其中领券中心只有一张图片，上面陈列了三件商品，频道广场有 2*3 六个站点，分别为母婴玩具、苏宁超市、苏宁家电、爆款手机、生活家电、苏宁国际，分类比较宽泛。猜你喜欢是按照每行 5 个

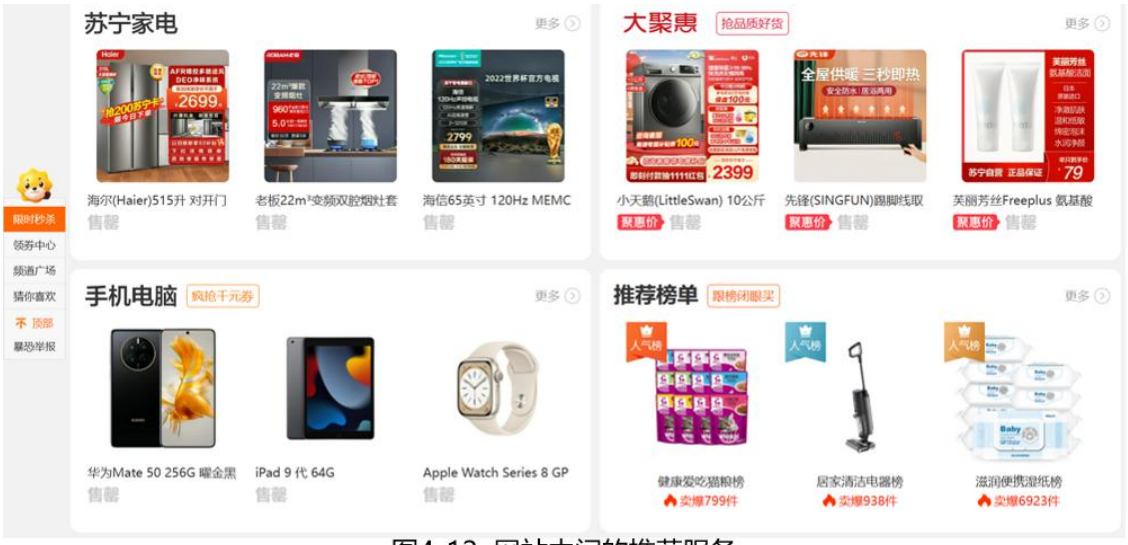

图4-12 网站中间的推荐服务

商品的方式向下陈列,商品的排列应该是根据主页的六个排序影响因子进行排列的,或者商家购买的推送服务。

图4-13 苏宁的推荐服务：频道广场

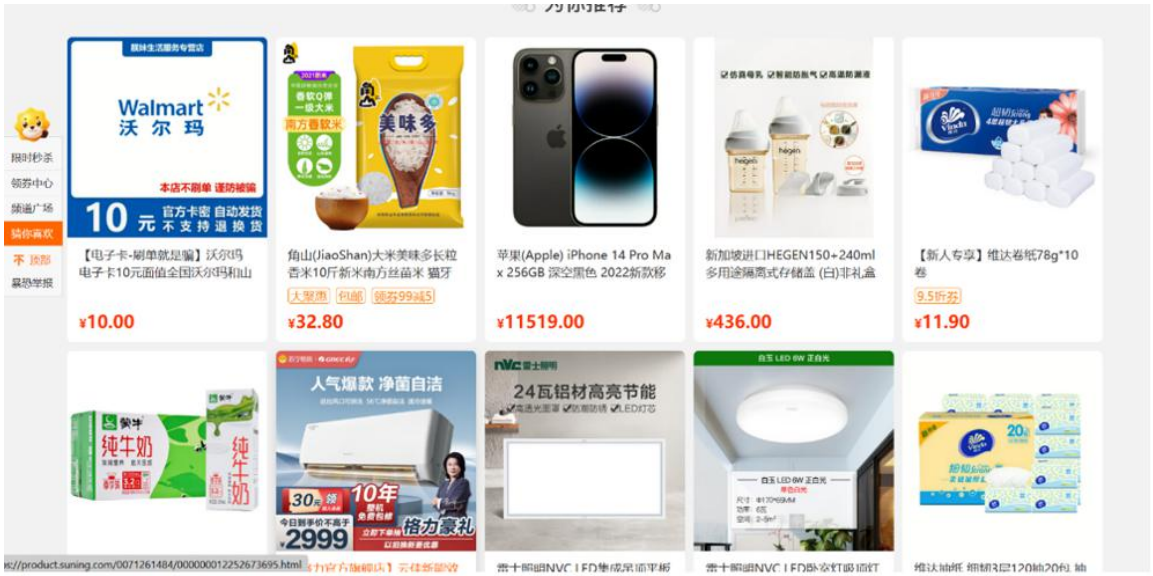

图4-14 苏宁的推荐服务：猜你喜欢

3.2 拿家具举例，从页上的个性组织服务

家具页面的推荐组织服务排列如图 4-15：

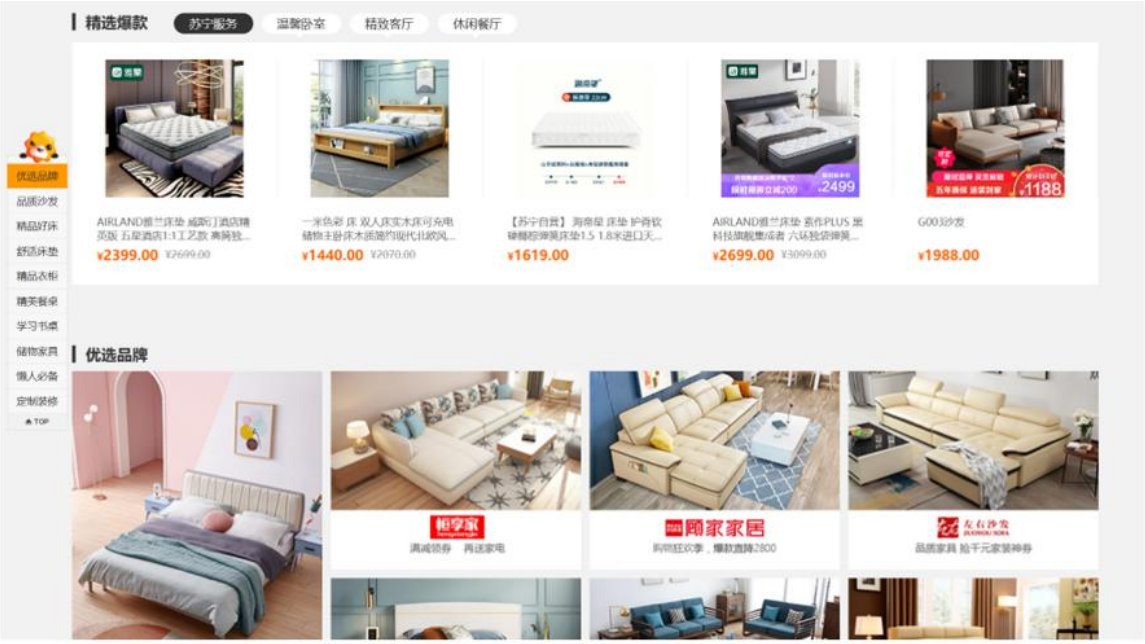

图4-15 商品家具从页的推荐服务

可以看到从上到下按“优选品牌”、“精品好床”、“舒适床垫”、“精品衣柜”、“精品餐桌”、“学习书桌”、“储物家具”、“懒人必备”、“定制装修”排列。其中优选品牌放置了七张图片，最左边一张竖版图片，右边排列了 2*3 六张图片，整体上一目了然，且图片大小适中，能够较好地帮助选择著名品牌。

从品质沙发到个性装修都采用了最左边一张竖版图片，内容是某商家，右边 2*5 十张方形图片，内容是具体商品的排列方式，不知道什么原因，金额显示不明。商品的排列按照六个排序因子排列。

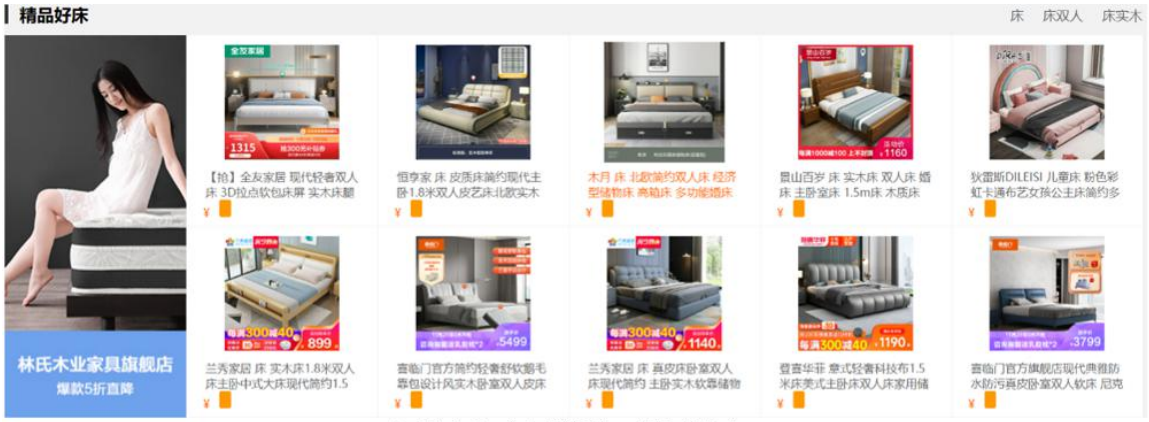

图4-16 家具推荐：精品好床

4.评价

4.1 主页的设计和安排

主页将主要的商品分类摆在了最显眼的位置，且没有采用全部列举的方式，用户浏览方便，图文并茂、内容较为详细。主页有一些动态图像广告，整体稍微有影响观感，网页信息组织方式多样化，分别按主题、产品、用户等方式分类，包含了所有应具备的必备事项。除了常见的搜索功能、联版广告、分类导航 icon 区以外，苏宁易购还在首屏放入几个主推模块作为主打产品。紧随其后的是场景化分类展示区，这个区域根据生活场景将频道与内容归类整理，方便用户浏览。其次则是直播及视频、热门市场入口、单品推荐区。这些大区域的区分方式注意了排版的统一性，利用纵向栅格使得上下对齐严格统一，横向栅格则使页面节奏统一，以达到提升浏览效率，层级更加清晰的目的。

在文字和图片排版上，苏宁主页采用了 2016 年横扫硅谷的新趋势 Complexion Reduction 风格，有几大特点：

标题字号加大、加粗。

图标简化。

黑白两色界面。

这一设计理念使得产品界面摆脱了无用的模块拼接式设计，走向了一条更加注重整体感、重视用户感受的道路。 CR 的设计风格使用大量留白，减少分割线等，来使界面更为清爽，此次首页也借鉴这种理念，尽量减少分割线的使用，利用留白和色块区分层级。旧版的首页由于放出了较多频道的 LOGO 字符使得页面字体较多，显得杂乱，新版使用统一的字体来避免出现体验满意度下降的情况。

4.2 组织结构

苏宁易购的组织结构为树状与网状结构的结合，主页分为若干栏目，每个栏目还含有子栏目，网页与网页之间可以相互跳转，比如图 4-17 中，选择空调后依然可以通过左上角的“分类”进入其他商品界面。所有网页都可以与主页链接（左上角有“返回主页”），网页与网页间也可以相互链接，条理清晰。

图4-17 苏宁首页的组织结构

首页的分类将网站所有内容组织成一颗由主页生发的树，用户浏览网站的过程是沿着树上下流动的过程，访问者可以根据路径清楚地知道自己所处板块的位置，不会迷路，而且随时可以达到自己喜欢的商品页面。

类目设置上根据用户的特点采用用户关注最高的 45 个类，这些类划分比较合适，即没有太宽泛也没有太狭隘，适合商品的检索和搜寻，也大体反映了网络资源内容的整体分布，一级类目的设置按照最符合用户认知的方式而非字顺划分，二级类目也按照检索频率确定，使得类目的排列顺序易于变动。比如空调类会把品牌列在上面，因为这是用户平常比较关注、搜索频率较高的点。

类目层次上设置不深，最多是 4 层，比如家具-客厅家具-沙发-实木沙发，因为分面组配法占比较大，很多商品上下从属逻辑不是很强。

类目注释上几乎没有详细说明，不过内容的本身也不太需要注释。

4.3 链接

有会员专享、精选爆款等多种链接方式，字体加粗加大，但没有变颜色，略显单调。大多数链接会产生一个新网页，而且在产生的新网页都有返回主页的链接，可以防止失误操作。

4.4 用户友好性

网站最下方有购物指南、售后服务、商家服务等，用户的使用比较方便，使用中的问题大体可以得到解决。

实验五

1.网站主题确定

1.1 分析的电子商务网站

京东

1.2 相关主题板块

品牌闪购

2.信息组织分析

2.1 信息资源类型

京东“品牌闪购”网页以及所包含的文字和图片类信息

2.2 信息组织方法

图 5-1 网站主页

首先，顾客在进入该网站后，可以在网站上部的索引中了解到：该网站展示了京东超市、服饰内衣、日用厨具、3C 数码以及更多中包含的母婴童装、家用电器、鞋靴箱包、运动户外、

钟表奢品、医药保健、珠宝配饰、家具建材、美妆护肤、品牌预告共 14 类，该分类体系主要以不同事物主题等设定相关类目，各个类别基本涵盖了人们日常生活购物需求的各方面，重点满足了使用用户注重生活、注重品质的新时代需求。而且每个类目的定名通俗简洁，汇集大众化的词汇，契合了广大用户层次差异化的需求。

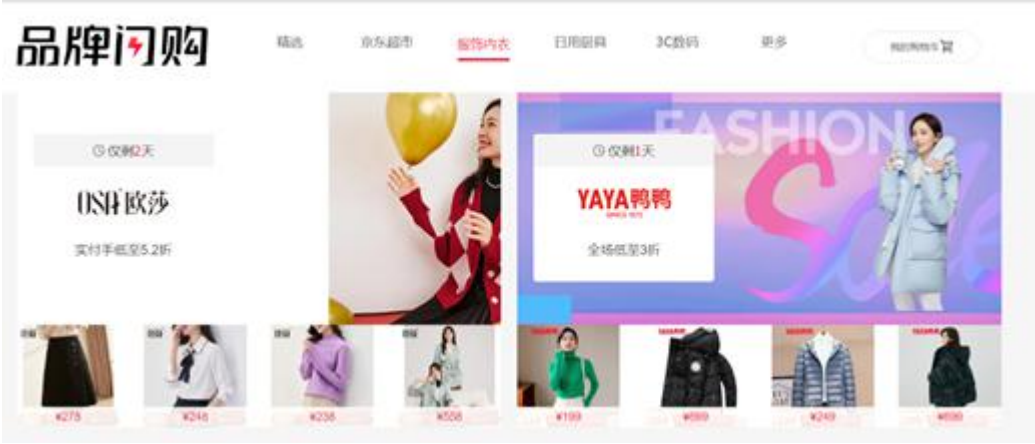

图 5-2 “服饰内衣”主页

在服饰内衣类目下，该网站设计了以品牌作为二级类目的方法，在页面下有各个品牌及其 4 个相关商品图片和对应价格展示，在每家品牌商标周围附有相关促销活动的信息。整体比较简洁美观，便于用户直接了解该品牌商品和促销信息，但是在该页面的品牌排序上则比较杂乱，没有明显的规律，当顾客想要寻找某一品牌时，没有相关索引提供帮助。

图 5-3 某一品牌主页

图 5-4 商品展示

图 5-5 商品分类导航

点击进入该品牌页面后，左方依旧是品牌商标，促销活动以及促销活动相关时间，右方则是相关图片展示，和主页的展示基本保持一致。下方则是品牌商品分类，排序方式，商品展示区域。

该平台提供的分类方式大致按以体系分类法为主，组配分类法为辅。在截取的两个品牌商品组配分类中皆是以男士女士作为一个分面，产品样式作为第二个分面。

在商品排序分面，平台提供了四种排序方式，分别为综合排序、销量、价格升序和降序。综合排序在一定程度上综合了销量和价格两个维度，对商品进行排序，综合排序的前四个商品同时也会展示在上一级的品牌页面的下方。当使用者选定某一分类，如选定“针织衫”时，平台会依据全部商品的相关描述与分类词进行匹配，剔除不满足条件的商品，将留下的商品按排序规则进行排序展示。

2.3 分析基于用户兴趣的组织方法

关于用户对信息内容的兴趣度表达，京东提供了多种方式。

一是用户可以将商品加入购物车。当用户将自己喜欢的、感兴趣的商品加入购物车后，平台会依据你所感兴趣的商品进行推荐。例如展示其他购买该商品的用户所购买的其他商品，依据你所感兴趣的商品通过相关算法推荐相关商品等。用户从

产生购买意向、到经历购买决策、直至最后下单的整个过程，在任何一个购物链路上的节点，推荐产品都能在一定程度上帮助用户决策。京东的推荐系统通过机器学习模型，结合知识图谱，挖掘商品间的关系，按用户场景，通过高维特征计算和海量召回，大规模排序模型，进行个性化推荐，提升排序效率，给与用户极致的购物体验。

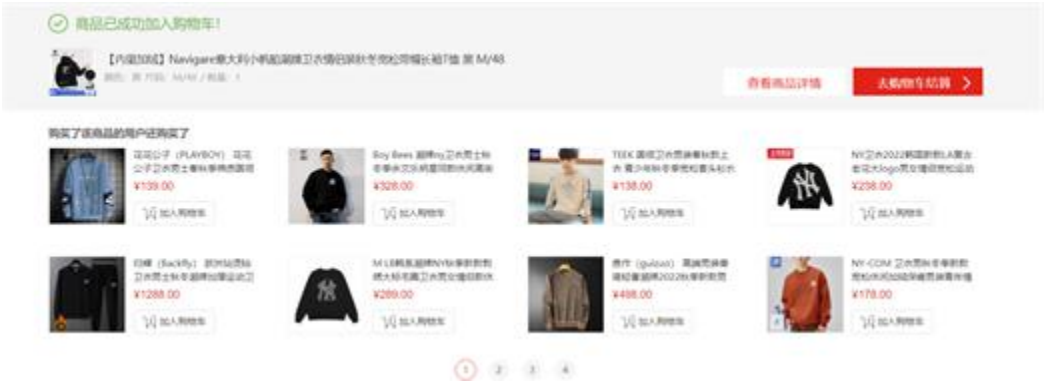

图 5-6 购物车

图 5-7 商品推荐

二是用户对商品或店铺进行关注。在关注商品上，平台为用户提供了这一功能，与加入购物车不同的是，它更多所容纳的是用户当前感兴趣却暂时没有购买意愿的商品，这些商品大多可能是因为用户当前预算不足、对价格不满意或者是购买不符合当前时期等，针对当前价格过高这一问题，平台还人性化地提供了降价提醒这一服务。在关注成功这一界面，平台设计“专辑”功能，其功能相当于建立收藏夹，将关注的商品进行分类整理。但是在笔者尝试该功能时存在不能创建专辑的情况，平台该功能有待优化。关注店铺的功能，旨在帮助用户收藏喜欢的或感兴趣的店铺，便于用户关注该店铺相关活动信息以及购买店铺商品。

图 5-7 商品关注

三是用户对商品的评价。用户在购买商品并收到商品后，可以对商品进行评价，评价分为好评、中评、差评，当用户在评价之后还可以对商品进行追评。在评价里，用户可以根据自身对产品的感受写出对商品的评价，同时还可以附上图片或视频。用户可以通过查看他人的评价来塑造对商品的印象，当评价中出现诸如商品很差劲的评论时，用户可能会放弃对该商品的购买，反之则提高对该商品的购买意愿。当用户认为该评价对自身有帮助时可以为评价进行点赞，高赞的评价也会优先展示在评价区。同时商家也可以根据评价来解答用户所遇到的问题，以及通过评价的反馈来改良自身的产品。

2.4 分析某商品的描述方法

图 5-8 商品描述页面

图 5-9 商品描述页面 2

信息描述是指根据信息组织和检索的需要，依据一定的规则和标准，对信息资源的主题内容、形式特征、物质形态等进行分析、选择、记录的活动。信息描述的结果是信息记录，也称元数据、Metadata。元数据是指用于帮助识别、描述和定位网络化的电子资源的结构化数据，通过它可以揭示各类电子文献的内容和其他特征以方便检索，能够提高信息的利用价值，其典型的操作环境是网络环境。

以该卫衣作为例子，“【内里加绒】Navigare 意大利小帆船潮牌卫衣情侣装秋冬宽松带帽长袖 T 恤黑 M/48” 在这段信息记录上，商家从品牌，规格，用途，样式等方面介绍了该商品的相关属性，中括号里强调了其加绒的特点，该描述基本包含商品的所有特点，便于顾客了解商品信息，根据自身想法产生购买需求。在左方的展示中包含了商品展示视频、模特演示视频、商品图片、商品吊牌信息图片等，区别于文字描述，商品图片和视频更加直观，便于顾客能从视觉上能感受到商品的魅力，从而产生购买兴趣。在右方的“品牌闪购”中包含了产品的闪购价、可用优惠券、促销方案以及相关时间限制。在下方则是商品相关服务、尺码、颜色、配送地址、付款方式等。在下方的商品介绍中则有着对商品更为详细的介绍，包含了商品名称、商品产地、适用季节、适用人群、工艺、风格等共 26 个商品属性。能有效为顾客展示该商品的特色。平台通过如上多种形势对商品进行个别化描述，使用户能识别该被组织的资源对象，通过记录商品的各种特征，如商品名称、商品产地、适用季节、适用人群、工艺、风格等，供用户对其使用价值进行判断，决定是否选择该商品；商品的描述信息记录是对商品全部特征的简练介绍，通过信息描述能有效实现对商品信息资源的管理。

3.信息组织评价

3.1 主页的设计与安排

整体简明大气，能突出其网页设计理念，排版简洁明了，品牌 logo 字符在商品展示时都占据着较为明显的地位，与其“品牌闪购”形象相符。各种折扣信息展示也能方便部分对促销信息敏感的用户。

3.2 组织结构

类目层次为三层，层次设置较为适中，类目定名通俗简洁，汇集大众化的词汇，没有专业性很强的用语，生活化气息重，因而能有效契合使用者层次差异的需求差异；该平台网站所建立的第一层次中京东超市分类，内容范围不太明确，既包含了食品烟酒，又包含了玩具厨具等，所包含类目较多，与其他有所重叠，导向不明确，缺乏规律。

3.3 链接

在链接方面，在关注商品并加入专辑这一功能上，存在无法创建专辑的 bug，其他地方暂时未发现问题。整体上基本有效且合理。

3.4 检索

在主页面未提供检索服务，无法精准搜索自己所需要的品牌商品，从而增加了用户在网站上搜寻商品的难度，影响使用的效率。

3.5 用户友好性

整体页面都十分简洁，整齐有序。以独具特色的品牌分类作为基础分类，能有效为在购物时更注重品牌的顾客提供服务。网页的相关操作也十分简单，在主页选择相关的商品类目分类后，便可以根据自己对品牌的喜爱来调选商家。相关折扣信息也在一定程度上吸引着消费者。

团队分工

在实验初期，由团队整体进行实验主题及方案的探讨，每个人分别提出相关的主题方案,最终由投票决定实验主题为“新时代高校大学生就业质量影响因素分析”。鉴于我们团队有五个人，所以我们决定实验四五一并完成。在分工之后，各成员分别开展实验，除了各自的任务之外，团队同时也会针对实验中所遇到的困难，一起进行讨论分析，相互协作贯通，共同完成实验报告。

具体团队分工

由漆银权负责实验一部分，周承桂负责实验二部分，李宗霖负责实验三部分，杨中昊负责实验四部分，何雷负责实验五以及团队分工部分，最后由小组成员共同审核修改整合报告。

个人心得部分

漆银权：

通过本次实验，我更进一步了解了中国知网的一些基本构造与网页布局，也更深入体验了中国知网的各种检索方式，对于之后的文献搜集能够提供一定的帮助。但是我也深刻认识到自己对于信息组织知识的学习还停留表面,在真正运用的时候,依旧感觉力不从心，不能够灵活处理，而本次实验便是充分的体现了这一点。

对待实验中需要进行信息组织方法的分析时，一开始只能较为基础的判断出该网页部分的信息组织方法为哪些,但是并不能系统的展示出该种信息组织各种特征。另外，对于有的信息组织方式只知道它的大类，并不能准确区分，比如对于主题组织法，并没有很好区分它是属于叙词法，还是属于单元词的方式。还需要通过进一步的了解，才能够严格区分它们的所属。

除此之外，还有元数据的分析，起初只知道元数据是信息描述的结果，是关于数据的数据，但是并不知道它的具体体现是怎么样，使用它对于文献描述的意义何在？因此，通过本次实验，我学习到了很多很多，也认识到自己的不足。只有自己亲自动手进行了实践，才会知道学习并不能只停留在知识表面。就好比，刚开始老

师讲的知识我基本都听懂了，但是当自己需要使用到的时候就显得苍白无力，唯有在实践中灵活运用各种方法，才能够更好理解并掌握它。

周承桂：

经过本次漫长而充实的实验课程，我收获颇丰，主要包含以下三个方面的内容。

第一，本次实验加深了我对本门课程知识的理解和运用，信息组织检索，不能仅停留在课本，必须亲身动手实践。在本学期的信息组织检索课程中，我们学习了很多基础理论知识，虽然课堂上有相关实例的讲解，但是毕竟我们没有自己动手去探索过，对于很多知识还是一头雾水。通过本次的实验课程，把理论知识和实际检索结合起来，完成一个个的实验内容，我对这些知识的认知便清晰许多。

第二，本次实验使我对于检索系统有了初步的认识，开始接触外文数据库，为以后的学习生涯打下基础。本次实验中我主要负责实验二的外文数据库实验，选取了 EI 检索系统。我结合课程理论知识，搜寻相关信息，逐步探索 EI 检索系统，从未知到认识到使用该检索系统，拓宽了我对检索系统的认识。

第三，本次实验使我深刻认识到在团队任务中做好团队沟通协作和有一名队长的重要性。在实验课程的开始阶段，我们团队每个人对于实验内容都比较茫然，不知道该如何入手，进行实验。由于我们团队并未设立队长，缺少一个人领导，指明工作方向，大家的交流也比较少，所以前中期我们的实验工作进展非常缓慢。直到实验后期，我们一起沟通讨论，分配好实验任务，快速推进实验，在工作中互相协作，最终完成了实验任务。这个过程使我认识到，一名队长对于一个团队是非常关键的。

李宗霖：

在本次实验课程中，我主要负责实验三——搜索引擎部分的内容。通过实验，我对搜索引擎的使用方法有了更深入地了解，知道了很多搜索引擎中提供的高级检索方法，对以后使用搜索引擎进行信息检索有很大的帮助。同时，在选择搜索引擎的过程中，对市面上提供的搜索引擎服务有了更全面的了解，知道了很多以前不曾听说过的搜索引擎，比如 360 搜索等，并对搜索引擎的原理及流程进行了学习了解。

在此基础上，分析了搜索引擎的信息组织方法与特点，将课程学习知识点与实际相联系，并对新型的网络信息组织方法有了深入了解。在信息描述方面，又通过论文学习与整理，细致地学习了课程中未接触的元数据的使用等方面的知识，对信息组织学科的内容了解更加全面。

最后，在信息组织的评价中，又对所学的第十章的内容进行了运用，但在实际操作中遇到了困难，最终选择用文献阅读的方式对 Yahoo 搜索引擎的信息组织进行评价。

通过本次实验，对课程所学知识有了全面的复习，实验三中涉及到大部分的所学知识。同时，在深入学习和完成实验的过程中，通过文献阅读，资料搜索，对实验内容及学科知识有了更深的了解。在完成个人实验部分后，又参与了其余部分的实验，对其它几个实验的思路、内容有了了解，对课程、学科有了更大的兴趣，弥补了在正式上课过程中的缺漏。同时，对文档编辑也有了心得，但仍存在问题，比如本次实验时间仓储中，对图片没有进行标号，也存在小的格式问题。

杨中昊：

在本学期的课程以及实验中，我对信息组织与检索有了初步的了解，知道了信息组织的作用，一些信息组织的方法，以及检索的方法。

在未学习本课程时，我也产生过这样的疑问，就是如果说动植物的分类可以根据 DNA 还有生物的特征进行分类，这些分类是有确凿的生物学原理的，是比较客观真实的，那么对于文献、图书以及网络资源这些涉及面广泛、内容繁杂、主题可能不统一的信息资源，由于主观性较强，该如何对其进行比较科学有效的组织。在本课程中，我认识到了有关文献内容组织的分类法、主题法，前者根据事物的性质将它们划分到不同的类目里面，给予不同事物不同的标号，并标号来索引，后者通过各种方法提取文献主要内容表示成主题词，然后通过主题词的组配来索引。在这其间为了保证分类的准确、灵活可可用性，还会使用空号、扩号、直接或相关索引、标题词、单元词、叙词语言等技术使得整体更加科学客观，让分类工作更成体系、有明确一点的标准，可以说这些分类方法很好地回答了我的疑问，也让我有了初步

建立分类体系的能力，在本课程作业中我们小组就设计了一套分类法，我们了解了分类的具体流程、查阅了相关文献，并讨论了一些分类中遇到的具体问题，比如分类应该用什么标准、哪些因素比较重要，最后也得到了让我比较满意的成果。

而对于网络资源的组织，当今最主要的问题是让计算机能够理解和处理信息，要实现这点就要用一些标准化的方法使信息得到有效组织，信息组织后不仅计算机可读，也要让人能够读懂。在本课程中我知道了像 MARC、XML、RDF、MP 等模型，这些方法能让我们在充斥着自然语言以及陈述不是很规范的网络上使用计算机去了解信息资源的内容以及组织信息。我还了解了检索模型以及检索结果的评价，向量空间模型、概率模型、潜在语义检索等，知道如何检验检索的效果以及如何找到文本中与检索最为相关的、排除一些具有迷惑性的、挖掘一些潜在性的、可能相关的信息。本次实验课程中，我们小组就对一些网站的信息组织方式进行了分析，而且设计了一些检索表达式，并评价检索的结果，让我们更加清楚了网站信息资源的布置方式、链接方式以及如何进行更有效的检索。

何雷：

经过本次实验课程的学习，我对信息组织与检索这一课程的了解更加深入了。本次实验我主要负责的是实验五，在实验开始初期，我们对课程内的部分概念了解还感到云里雾里，对于实验的要求感到疑惑，随着时间的变化，我们在老师的帮助以及组内的讨论下才最终决定了前面的实验主题，在确定分工之后，我们才徐徐开展各自的内容。

在实验五中，是电子商务的信息组织实验。经过一番思虑，我最终选取了京东购物网站中的“品牌闪购”模块。与京东其他模块不同的是，它在最基本的大类分类之下的第二级类目是按品牌进行分类，这份“独特性”和其“品牌闪购”的理念也是十分相应。在撰写实验报告中，也是遇到了很多困难，对课程的了解不够深入，因此我尝试去查找相关文献，借鉴学习更多与实验课程相关的知识，在深入通读相关文献之后，对于怎样分析某商品的描述方法以及基于用户兴趣的组织方法这一方面，有了更为细致的了解。学习是一个循序渐进的过程，在初步学习之后，在运用

之中发现自身的不足，然后再通过学习去弥补自身的不足，再将知识运用到实践之中。经过本次实验，我也更加了解到自身的不足以及自身知识的浅薄，在该课程上的学习还不够充分。

本次实验所带来的经验不仅仅用于现在，更能启迪未来。我们是大数据专业，未来会遇见各种各样的新兴事物，注定会有着我们所空白、不了解的领域，学习是为了消除这种空白，更深入了解这一领域，从而创造更好的未来。