**调研报告**

**现在市场上有许多各式各样的电子商务平台，外卖、服装、旅游，电子商务已经涵盖了我们生活的方方面面。基于上述环境，我们小组同样计划编写一个水果买卖行业的电商平台，因此在发行之前为了更好的进入市场于是便进行了以下调研报告。这份报告主要是基于我们开发软件所使用的mule平台搭建后台技术与其他电商软件语言的对比以及推荐算法的选择。**

1.**行业平台搭建技术研究及我们使用mule的优势**

目前网络上的电商平台所使用的语言主要有以下几种：1.php 2.jsp 3.asp.net 4.java四种。1.php支持使用平台丰富,lamp(linux+apache+mysql+php)可以说是PHP支持的最佳平台了。免费，开源，安全，开发成本低，速度快，负载强。同时也可以使用Apache+PHP+MySQL在windows上跑。目前国内的主机基本上都支持。非常适用于构建中小型的电子商务，开发快，易于维护，开发出来的WEB速度也不逊色。目前使用php开发的一些电子商务网站有：淘宝、shopex、ecshop、shopnc等。2.jsp的跨平台性，有标准的架构，能够支持高度复杂的基于Web的大型应用。但是其开发成本较高，java+Oracle解决超大型运用是首选。目前使用jsp开发的一些电子商务网站：亚马逊、拍拍、有啊、易趣、橡果国际等。3. asp.net的快速开发比较强大，强大的IDE开发工具与调试功能，使得快速开发成为可能。但是.net不能跨平台并且相对来说安全性较差，不过事在人为。但.net的CLR使其效率可能有所流失。另外，由于ASP.NET本身支持企业技术，例如消息队列、事务、SNMP和Web服务，因而可以很容易地开发具有高度可缩放性的强大的应用程序。使用asp.net开发的一些电子商务网站：京东、当当、新蛋、携程、凡客、shopwe等。4.java算是我们最熟悉的语言了，高度的跨平台性使得他一直被广泛使用，实际上jsp也是基于java的。近来很多电商平台都在逐渐更换自己平台的语言变为java其实是看中了java的稳定性，它为高并发性访问提供了可能而不需要大量增加服务器的数量。从以上语言的介绍可以看出电商平台选择语言时的影响因素1.易开发性 2.跨平台性 3.快速开发性 4.稳定性。易开发性：几乎所有的广泛使用的电商程序开发语言都是有着良好封装，易于开发者使用的。借助良好的封装工具开发者能够更容易完成程序。同时，易开发性还表现在开发者对语言的熟悉程度，每个开发团队擅长的语言都是不一样的，重新学习一门语言的价值可能会高于语言本身劣势带来的损失，如此为了一些语言上的优势而重新学习一门不熟悉的语言常常得不偿失，特别是对于一些小的新的团队而言，例如我们。对语言的熟悉的常常决定了一开始语言的选择。2.跨平台性。电商平台现在往往在多个不同的平台上使用虽然主流的开发方式仍然是在windows系统上进行但是跨平台性仍然是一个很重要的指标，能够跨平台的语言对于开发的持续进行是很有利的，当然这并非是最主要的考虑因素，它只能够为选择的语言加一些分数而已，毕竟比起其他的因素跨平台性是比较容易克服的特性。3.快速开发性。电商平台的一大特点就是客户需求变更迅速，为了不断推出新的版本以及维护，快速开发性能是必须的。可以看得出来，具有快速开发性能的语言能帮助开发者在市场上确立优势并且及时调整战略，更新软件，避免因为决策失误或者市场变化而无法及时调整而出现大的问题。4.稳定性。电商在占领了一定的市场以后无可避免的问题就是大并发量同时访问，语言的稳定性可以说是保障用户体验的至关重要的环节，通过稳定性高的语言开发者可以节省服务器的数量同时也避免了一些使用过多服务器而带来的问题。在电商发展到规模越大的后期，稳定性的重要性就越能体现出来，这也是为什么一些发展的好的平台例如淘宝等开始或者已经将其系统换为稳定性更高的java语言的原因。

相对与外界普遍使用的以上四种语言，我们经过挑选选择了mule这个框架来构建我们的后台。Mul e 是世界上使用最为广泛的开源企业服务总线和集成平台之一，它是为支持不同系统和服务之间的高性能、多协议业务而设计的。Mul e 提供了一个分布式的基于事件的架构，其功能是在两个分离的应用程序和服务之间传递数据。Mul e 支持同步、异步和请求响应等事件处理和传输机制，如JMS、HTTP 等。它能很容易地嵌入到任何应用框架中，并且明确支持Spr i ng 框架。

Mul e 可安装在不同的操作系统上，如：Li nux、Wi ndows、Mac OSX等，也可以支持很多不同的J2EE 容器，如：ApacheTomcat、WebLogi c、WebSpher e、Jboss、Or acl e、Jet t y 等。同时Mul e 本身也集成了众多的服务功能，如：消息代理、业务流程编排、路由、传输转化等等。在其广泛的平台兼容和丰富的集成服务的基础上，Mul e 可以提供强大的功能满足集成上层不同的业务需求，如：业务应用、丰富的客户端和门户、传统的B2B应用、数据库应用等等。同时我们使用java来搭建我们的前端（事实上mule也是基于java的），使用mySQL作为数据库支撑，这三者的配合很好的实现了快速开发以及稳定性，使得未来软件的持续使用以及进一步开发成为可能。同时因为我们团队的软件编写成员对于这三者都非常熟悉，他也实现了易开发性，同时跨平台性可以通过更换数据库软件来实现。可以说这样的框架选择对于我们小组来说是十分完美的。

**2.搜索算法调研及选择**

对于推荐算法目前来说我们小组所设想的是使用简单的加权值方式来更有针对性的推荐，目前来说根据相关的资料显示，一种对推荐算法的分类方式将推荐算法分为了三类： 基于内容的推荐算法、协同过滤推荐算法和基于知识的推荐算法。  
基于内容的推荐算法，原理是用户喜欢和自己关注过的Item在内容上类似的Item，比如你看了哈利波特I，基于内容的推荐算法发现哈利波特II-VI，与你以前观看的在内容上面（共有很多关键词）有很大关联性，就把后者推荐给你，这种方法可以避免Item的冷启动问题（冷启动：如果一个Item从没有被关注过，其他推荐算法则很少会去推荐，但是基于内容的推荐算法可以分析Item之间的关系，实现推荐），弊端在于推荐的Item可能会重复，典型的就是新闻推荐，如果你看了一则关于MH370的新闻，很可能推荐的新闻和你浏览过的，内容一致；另外一个弊端则是对于一些多媒体的推荐（比如音乐、电影、图片等)由于很难提内容特征，则很难进行推荐，一种解决方式则是人工给这些Item打标签。  
协同过滤算法，原理是用户喜欢那些具有相似兴趣的用户喜欢过的商品，比如你的朋友喜欢电影哈利波特I，那么就会推荐给你，这是最简单的基于用户的协同过滤算法（user-based collaboratIve filtering），还有一种是基于Item的协同过滤算法（item-based collaborative filtering），这两种方法都是将用户的所有数据读入到内存中进行运算的，因此成为Memory-based Collaborative Filtering，另一种则是Model-based collaborative filtering，包括Aspect Model，pLSA，LDA，聚类，SVD，Matrix Factorization等，这种方法训练过程比较长，但是训练完成后，推荐过程比较快。  
最后一种方法是基于知识的推荐算法，也有人将这种方法归为基于内容的推荐，这种方法比较典型的是构建领域本体，或者是建立一定的规则，进行推荐。（作者：张海东  
链接：http://www.zhihu.com/question/20326697/answer/58148605  
来源：知乎  
著作权归作者所有。商业转载请联系作者获得授权，非商业转载请注明出处。）

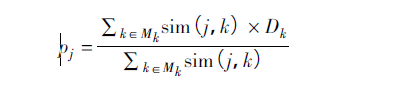
可以看到对于我们平台来说作为一个水果售卖平台我们更关注的是用户个人的兴趣但是从短期来说我们需要一种能够较快地通过其他相似用户获得用户最可能喜欢的水果以及说过周边商品的信息，为此我们查阅了相关资料寻找到一种基于用户兴趣特征提取的推荐算法，此算法的核心为构建用户兴趣度矩阵借此分析用户的兴趣并且对于用户兴趣转移的情况下仍具有较高精确度，以下是其构建矩阵以及产生具体推荐集的方式：

设用户u 在商品空间集Uu，j中未浏览过的商品集合用Nu表示，浏览过的商品集合用Iu表示，则Nu = Uu，j － Iu，对任意商品p∈Nu，预测用户对商品p 的兴趣度方法如下:a) 在商品集Iu中寻找对商品i 和j 览过的所有用户的兴趣度值，利用传统的相似度度量方法计算商品i 和j 之间的相似性。

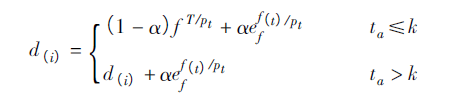
b) 选择相似性较高的k 个商品构成商品相似度集合Mk ={ I1，I2，…，I k} ，其中pMp，并使sim( p，I1) ＞ sim( p，I2) ，… ＞sim( p，I k) 。

c) 本文利用文献［10］提出的方法预测用户对未浏览过的

商品的兴趣度:

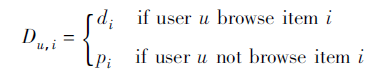


通过上述方法处理后，结合式

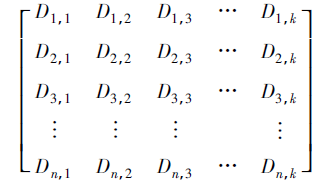


（于原文前文中已证）

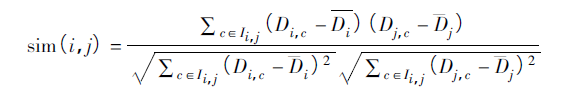
得到用户u 对任意商品i 的兴趣度为



在此基础上构建n × k 的用户u 对商品j 的兴趣度矩阵，n行代表n 个用户，k 列代表选择的k 个商品，Di，j表示用户u 对商品k 的兴趣度值，用户兴趣度矩阵如下:



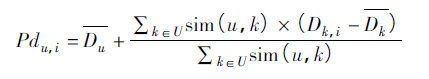
为了产生精确推荐结果，利用目标用户的邻居用户对商品的兴趣度值预测目标用户对未浏览商品的兴趣度值，并选择兴趣度值较高的top-N 个商品推荐给目标用户。根据建立的用户对商品的兴趣度矩阵，利用Pearson 相关度计量方法寻找与目标用户兴趣相似的邻居集合，在此基础上预测目标用户对未浏览过的商品的兴趣度。相似度度量方法如下:



其中: Ii，j表示用户i 和j 共同感兴趣的商品集合，Di，c表示用户i对商品c 的兴趣度，Di表示用户i 对所有商品兴趣度的平均值。

于目标用户u，在整个用户空间中寻找与其兴趣相似的用户集合U = { u1，u2，…，uk} ，使得uU，并且sim( u，u1) ＞sim( u，u2) ＞… ＞ sim( u，uk) 。

根据用户集合推测目标用户对未浏览商品的兴趣度值方法如下:

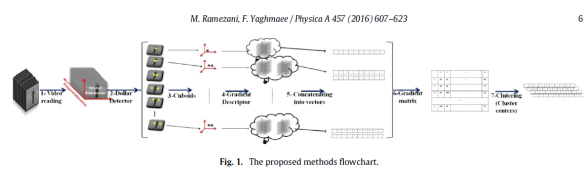


其中: Du和Dk表示目标用户u 和邻居用户k 对商品的平均兴趣度， sim( u，k) 表示目标用户u 与邻居用户k 的兴趣相似度，Dk，i表示邻居用户k 对商品i 的兴趣度。

除此之外我们还找到了一种基于人类行为的推荐算法，可以考虑使用于老用户的推荐。以下是一部分关于其算法具体使用的描述，由于是英文资料所以可能由于语言能力问题未能非常完整地截取其所有主要部分：

Here, in order to recommend some videos to a user, human action in each video must be modeled. To this end, somefeature point of actions are extracted that their motion directions and scale are shown in some vectors. These vectors areused in different ways to represent each video. Here, beside a novel low complex representation method, a fuzzy distance

measure is presented to calculate the dissimilarity of two videos.

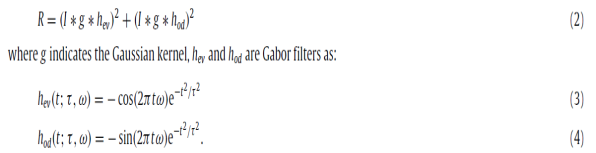


3.1. Retrieving the video

Today, retrieval systems are used to find the most relevant real world videos to a query one based on the contents. Insuch systems, time efficiency and scalability can be considered as their main challenges. In our method, some local featurepoints of each video are initially extracted. To this end, Dollar detector [40] which is sensitive to repeating motion patterns is used. In before methods like Refs. [34,31], gradient information of feature points are shown in some vectors and they are used for representing the action in video.

Fig. 1 shows the proposed methods’ main steps and the initial three steps appertain to Dollar method procedure. In this method, distinct linear filters are applied on each video sequence in order to calculate a response functions for points (as Fig. 1 (steps 1 and 2)). One of these linear filters is a 2D Gaussian kernel filter in the spatial dimensions and the other one is

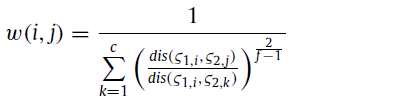
a quadrature pair of 1D Gabor filter on the temporal axis. The 2D Gaussian kernel filter is applied as follows:



After detecting feature points, a spatio-temporal cuboid is considered around each STIP (as Fig. 1(3)). Then, a gradient descriptor is applied on each cuboid to calculate the point’s main motion direction and scale (as Fig. 1(4)). Results of the gradient descriptor on each axis are concatenated into one vector namely gradient vector of the STIP (as Fig. 1(5)). In our method, no dimension reduction algorithm is used to avoid missing any motion information of the STIPs. In 6’th step, a matrix of the whole gradient vectors is considered for each video, namely gradient matrix. Note that, the equal mean test is applied on gradient matrix of each video in order to analyze similar patterns in each dimension of vectors. Here, dimensions that vectors have similar pattern in them can be represented by their mean and the other dimensions are grouped based on the main motion directions. To this end, whole video vectors are grouped and their centers are supposed as the main vectors to represent the action. Furthermore, to group the initial vectors and calculating the representation vectors,

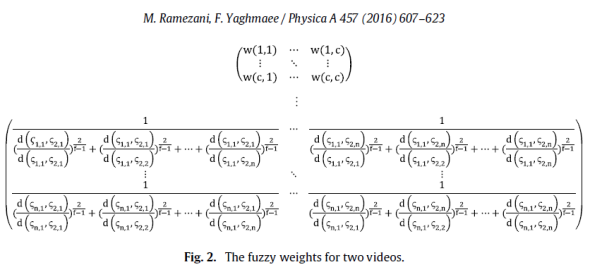
k-means and fuzzy c-means clustering algorithm are applied on the gradient matrix (as Fig. 1(7)). After representing videos, they are ranked based on their likeness to the query one which a dissimilarity measure is used for. This dissimilarity for two videos relies on fuzzy weights that are calculated based on the distance between the vector pairs of those videos. These fuzzy weights between two videos indicate the impact of corresponding vector pairs of those videos on final distance. Suppose that, vm indicates the m’th video of the dataset and vm has n interest points that are shown by {ξ1, ξ2, . . . , ξn}. After applying the clustering algorithm, c cluster centers are achieved for representing the m’th video. These representation

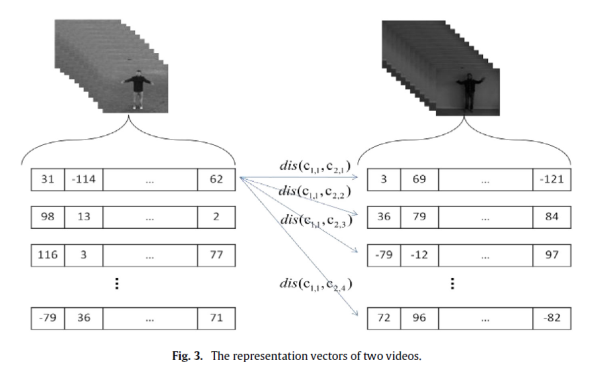
vectors are shown by ςi,1, ςi,2, . . . , ςi,c. Suppose there are two videos, namely υ1 and υ2, the fuzzy weight of a pair of the videos’ vector is shown by w(i, j), where i indicates the vector number of the first video (or its cluster number) and j indicates the vector number of the second video (or its cluster number). The weight between i’th vector of v1 and j’th vector of v2 is calculated as follows:



where, dis （ς1,i, ς2,j）indicates the Euclidean distance between the i’th vector of v1 and j’th vector of v2 and f is the fuzzy constant. Note that, the weights for two videos are stored in a c × c matrix that each cell indicates the weight of a pair of vectors (Fig. 2).

On the other hand, the weight matrix can be supposed as a binary matrix by stacking up the minimum distances between the vectors. As shown in Fig. 3, in a binary matrix, if the first vector of v1 has the least distance to the third vector of v2, then dis（ς1,1, ς2,1）, dis（ς1,1, ς2,2）and dis（ς1,1, ς2,4）are ignored and they have no effect on the calculation of the final distancebetween v1 and v2. In other words, while dis（ς1,1, ς2,1）may be less than dis（ς1,2, ς2,1）, dis（ς1,3, ς2,1）and dis（ς1,4, ς2,1）,its effect on the similarity of those videos would be ignored. In fact, for calculating the dissimilarity of v1 and v2 by a binary



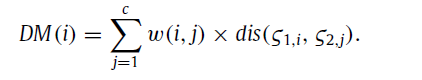


weight matrix, minimum distance between each vector i from v1 and the vectors of v2 will be considered and can be shownas follows:

C:\Users\admin\AppData\Roaming\Tencent\Users\381003545\QQ\WinTemp\RichOle\ZXM@_M66@{DX`X70~{~V1AX.png

where, DM (i) is the minimum distance between the i’th vector of v1 and vectors of v2 and c is the number of vectors(clusters).

On the other hand, for calculating the dissimilarity of two videos by enumerating the fuzzy weights, DM (i) can be calculated as follows:



Based on the computed DM (i), the final dissimilarity of those videos is calculated as follows:

C:\Users\admin\AppData\Roaming\Tencent\Users\381003545\QQ\WinTemp\RichOle\CPXAR5RS{E6H[GK78VH~}EU.png

where,DM (i) indicates the distance factor between the i’th vector of v1 and the whole final vectors of v2 and c is the numberof vectors (clusters).

以上是我们希望将来能够使用于我们软件的两种推荐算法，目前来说由于工期原因很遗憾我们的第一代产品只能借鉴这两种算法尽量将各个商品推荐时候的权值设计的尽量合理等待日后将其完善。

**3.结束语**

通过这次对市场的调研我们小组找到了合适自己搭建软件的平台，并且通过对比找到了我们的优势所在更加确立了我们完成一个水果导购平台的信心，同时我们还找到了未来可以帮助我们做的更好的推荐算法，这样我们就可以不限于即将产生的初代产品而将目光看向更远的2.0、3.0代，对于我们未来的发展是很有利的。

4.参考资料

\*A novel video recommendation system based on efficient retrieval of human actions Mohsen Ramezani, Farzin Yaghmaee∗

Department of Electrical and Computer Engineering, Semnan University, Semnan, Iran

\*基于用户兴趣特征提取的推荐算法研究\*

刘枚莲，刘同存，李小龙

( 桂林电子科技大学，广西桂林541004)

\*开源中间件Mule 的研究

Research of Open-source Middleware Mule

彭政

Peng Zheng

(电子科技大学中山学院，广东中山528400)

(Zhongshan Institute, University of Electronic Science and Technology of China, Guangdong Zhongshan 528400)