基于时间序列分析的 SNS 开放平台应用研究

裴珊珊[1,2]　叶小梁[1]

[1]（中国科学院文献情报中心 北京 100190）

[2]（中国科学院研究生院　北京 100049）

【摘要】介绍 SNS 开放平台及其应用相关概念，分析开放平台应用现状和运营模式，基于国内此类软件的统计数据，采用 DTW 算法度量变长时序数据的形状相似性，使用一维序列变换改进 K 中心点聚类的簇中心计算方法，并对数据进行时间序列聚类挖掘，最后使用产品生命周期理论分析聚类结果。

【关键词】SNS 开放平台 时间序列 聚类分析

【分类号】TP393

Time Series Analysis of SNS Open Platform Applications

Pei Shanshan[1,2] Ye Xiaoliang[1]

[1] (National Science Library, Chinese Academy of Sciences, Beijing 100190, China)

[2](Graduate University of Chinese Academy of Sciences, Beijing 100049, China)

【Abstract】Firstly, this paper introduces concepts of SNS open platforms and their applications, concludes the current status of these platforms and applications. Then, it uses a series transformation method to calculate the cluster centroids of K-Mediods clustering algorithm. The clustering experiment is conducted with Dynamic Time Warping algorithm as the distance measure of series with various lengths. Finally, the paper analyzes the experiment results with product life cycle theory.

【Keywords】SNS　open platform　time Series　clustering

1　引言

近年来，Web 2.0 得到了充足的发展，在以人为中心的网络环境下，Facebook，MySpace 等 SNS 聚集了巨大的用户群，也积累了大量的用户信息和用户资源。与用户量伴随的是与之匹配的市场，为有效利用 SNS 的用户资源、寻找基于 SNS 用户群的商业利益、提供更丰富的用户交互服务，一些 SNS 开始提供可获取有限用户数据的开放平台，使得第三方开发者可基于平台开发供 SNS 用户使用的应用软件。本文将基于国内主流 SNS（校内网、51、聚友等）的开放 API 应用情况，以定量统计和聚类分析的方法对 SNS 开放平台应用进行探索研究。

2　SNS 开放平台应用现状

2.1 相关概念

（1）社会网络网站

Social Network Sites，简称 SNS，是提供具有现实或潜在关系的用户进行信息交互、分享和社会化交流的在线社区。典型的 SNS 包括 Facebook、MySpace、Hi5 以及国内的校内等等。

（2）SNS 开放平台

SNS 开放平台是提供使用 SNS 信息资源和服务的外部程序接口的软件系统。基于 SNS 开放平台的应用程序接口（API），第三方应用开发者可以将 SNS 的资源和服务集成到自己的应用软件中去。

（3）SNS 应用程序

SNS 应用程序是针对 SNS 开放平台定制开放的小规模软件程序，提供休闲娱乐、学习交流等丰富交互应用，并成为 SNS 用户个性化定制的应用组件。SNS 应用程序包括由 SNS 自己开发的应用程序和由第三方应用开发者开发的应用程序。典型 SNS 应用程序如移动接入、在线数据库服务、多媒体应用、社会化游戏（Social Game）等。

2.2 SNS 开放平台应用统计

自 2007 年 5 月 Facebook 宣布推出开放平台以来，基于 Facebook 平台的应用程序呈现出爆炸式的增长。2008 年，搜狐、聚友(MySpace.cn)、校内网、51.com 等多家国内 SNS 相继推出自己的开放平台并已取得良好效益。

表 1 国内外主流 SNS 开放平台数据对比

平台	应用数量	开放时间	平台标准	活跃用户（百万）
Facebook	57469	2007.5	Facebook F8	175
MySpace	7786	2008.2	Open Social	120
校内网	1176	2008.7	Open Social	22
51.com	270	2008.8	Open Social	31.5
聚友中国	268	2008.1	Open Social	N/A
搜狐博客	916	2008.1	Universal Widget API	N/A

数据来源： Facebook.com, MySpace.com, Xiaonei.com, 51.com, Sohu.com, Appdata.com, Comscore.com, www.appleaf.com, 截至 2009 年 3 月。

表 1 显示了国内外主要 SNS 开放平台的相关统计数据，其中 Facebook 和 MySpace 是国外用户量最多的 SNS，而校内网、51.com 等则在国内市场领先，它们的相关数据能够反映整个行业的开放平台发展水平。

对表 1 中 4 家 SNS 开放平台应用程序的监测表明，在开放平台推出初期，应用程序数量出现了快速的增长，但由于开发团队较少、平台审核等多种因素，国内开放平台应用进入了平稳增长期，如图 1 所示：

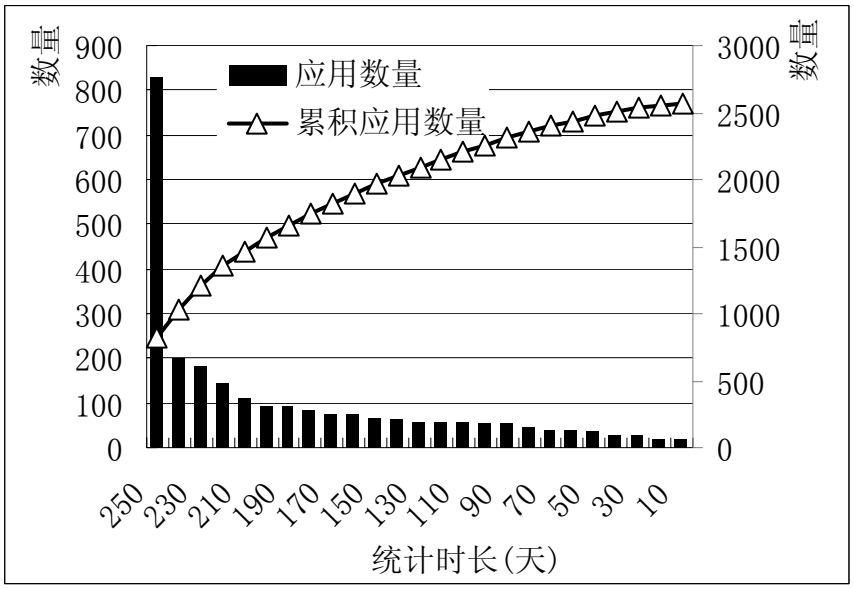


图 1 国内 SNS 开放平台应用的统计时长分布

数据来源： Xiaonei.com, 51.com, MySpace.cn, Sohu.com, www.appleaf.com, 截至 2009 年 3 月。

2.3 SNS 开放平台应用的开发与运营模式

SNS 开放平台应用的基本开发与运营模式是：SNS 依托庞大的用户信息资源建立开放平台，第三方开发者或 SNS 自身通过开放平台将用户信息集成到应用程序中去，并发布到自有服务器或 SNS 平台上，到达终端用户，如图 2 所示：

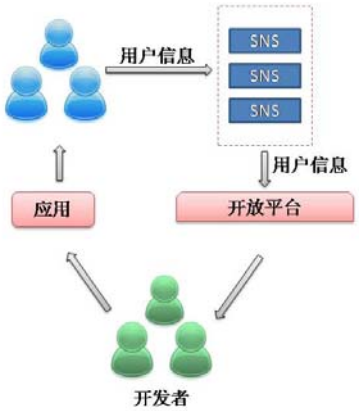


图 2 SNS 开放平台应用的基本开发与运营模式

3 SNS 开放平台应用的时序聚类分析

利用 SNS 开放平台应用程序的统计数据，我们可以对此类应用的生存模式和产品生命周期进行深入分析。特别地，对于具有时间特征的软件统计数据来说，时间序列挖掘的方法能够帮助发现 SNS 开放平台应用程序的规律、趋势和突变等。

在本部分，我们利用聚类分析的方法，尝试将应用的时序统计数据划分为若干模式，并在下一节使用产品生命周期理论进行分析。

3.1 数据收集与筛选

（1）数据收集

本文选取了校内网、51 网、聚友网、搜狐 4 家 SNS 的开放平台应用作为研究的数据来源，统计的数据指标包括：

① 应用安装用户量：SNS 用户中安装该应用的数量；

② 应用活跃用户量：安装该应用的用户中每天使用了该应用的数量。

（2）数据筛选

为保证数据分析的有效性和一致性，我们使用数据较为完整的校内网和 51 网开放平台上的应用作为分析对象。

时间序列分析要求被分析的对象具有一定的时间跨度，我们选取了至少具有 140 天统计数据、活跃用户数量不少于 100 人的应用软件作为最终分析的样本。

3.2 数据变换与归约

数据变换与归约将数据处理成一种适合挖掘的形式。本文涉及到的变换归约包括噪声平滑和规范化。

（1）噪声平滑

噪声平滑的目的是去除数据中的噪声。本文中的时间序列数据在多种情况下可能形成噪声数据点，如服务器瘫痪、网络失效等。

给定序列$S=\{d_1,d_2,...,d_i,...,d_L\}$，变量$d_i$可以看做时间点 i 的函数，数据变换的要求是不改变序列数据与函数趋势的基本匹配程度。

我们采用 n 阶加权移动平均来平滑时间序列，移动平均具有消除数据集中的变差的特性，因此能够消除不必要的波动。

$$d_i'=\frac{1}{\sum_{j=1}^{i+n-1}w(j)}\sum_{j=i}^{i+n-1}d_j*w(j)$$
$$w(j)=b-a*(j-i-n/2)^2 \quad (1)$$

加权移动平均的方法通过给予中心数据点较大的权重来抵消光滑带来的负面影响。

（2）数据规范化

开放平台应用的统计数据在数值上差异巨大，其中，最大的安装量达到了百万用户以上，而普通应用的用户量仅有数百，因此，在比较序列数据之前，需要将序列数值规范化到特定区间。规范化可以防止具有较大初始值域的样本与具有较小初始值域的样本相比距离过大。

经典的规范化技术包括最小-最大规范化、z-score 规范化、小数定标等方法。在本文中，我们使用最小-最大规范法来对数据进行线性变换：

$$v'=\frac{v-\min_{d_i}}{\max_{d_i}-\min_{d_i}}(new\_\max_{d_i}-new\_\min_{d_i})+new\_\min_{d_i} \quad (2)$$

其中，$\max_{d_i}$和$\min_{d_i}$分别为原始序列中数据的最大值和最小值，此方法会将序列数据映射到区间$[new_\min_{d_i},new_\max_{d_i}]$，默认情况下，设置$new_\min_{d_i}=0,new_\max_{d_i}=1$。

3.3 基于 FastDTW 的不等长序列距离度量

对数据进行聚类分析的一个基本条件是确定样本的相似度或距离度量方法，距离较小或相似度较高的样本最终被聚集到同一个簇的概率较大。在本文中，由于分析的对象是软件产品的成长规律，时间序列数据所形成曲线的形状和趋势更加重要。

传统的数据相似度度量或距离度量方法大多基于样本特征向量的相似度（如 Cosine 夹角）计算或直方图求交等基本方法，在时间序列相似度度量方面，大多也关注于序列变换降维、序列移步比较、子序列匹配等，但大多都没有很好地解决序列的形状匹配问题，特别是样本在趋势类似但偏移较大的情况下，传统方法的效果受到制约。

（1）距离度量函数的基本要求

考察应用软件相关时序统计数据的变化模式相似性，需要度量函数满足以下约束：

① 形状的相似性不受序列峰值区间所在具体位置影响；

② 形状的相似性不受到序列的长度影响；

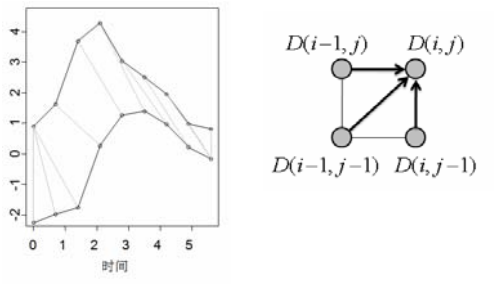
（2）序列距离度量方法

我们将 FastDTW 算法[3]引入为聚类采用的序列样本距离度量方法，FastDTW 是 DTW（Dynamic Time Warping）[4]算法的一种改进变形，被广泛应用于语音识别等领域。

DTW 算法可用于衡量在时域或者速度上不一致的序列的相似度。它通过在给定约束下寻找两个序列的最优匹配，在时域上实现序列的非线性缩放，从而获得序列间的距离或者相似度。

算法采用动态规划来逐步取得局部最优，给定序列 S_1, S_2，长度分别为 m, n，DTW 将序列的首尾数据点对齐，定义输出的距离函数 $D(m,n)$ 为两序列间的距离。

需要找到一条路径 $\{(p_1,q_1),(p_2,q_2),...,(p_k,q_k)\}$，使得点对点间的距离之和最小，并且 $(p_1,q_1)=(1,1),(p_k,q_k)=(m,n)$。



(a) (b)

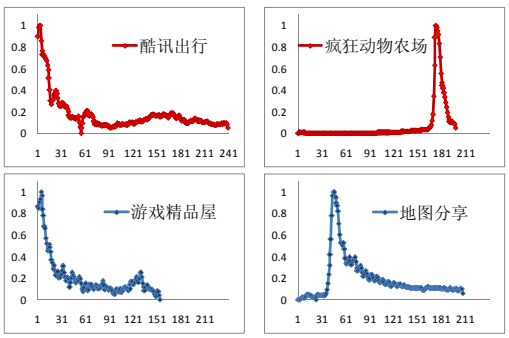
图 3 Dynamic Time Warping：(a)两条时间序列曲线的对齐；(b)路径的局部最优选择。[1]

于是，我们求取并记录局部最优的路径和点对。

$$D(i,j)=d(i,j)+\min\{D(i-1,j),D(i,j-1),D(i-1,j-1)\} \quad (3)$$

其中 $d(i,j)$ 为两点之间的距离度量。

图 4 给出了两个本文中不等长序列的距离度量实例，左右侧各包括第一行序列样本及与其距离最小的三个不等长序列样本。我们可以看到本文使用的距离度量算法可以有效地找出形状趋势相似的时间序列样本。特别地，对于右侧的实例，在总体变化类似的情况下，距离度量的算法对序列样本中峰值出现的位置较为鲁棒。注意数据均被最小-最大规范化到[0,1]区间内。



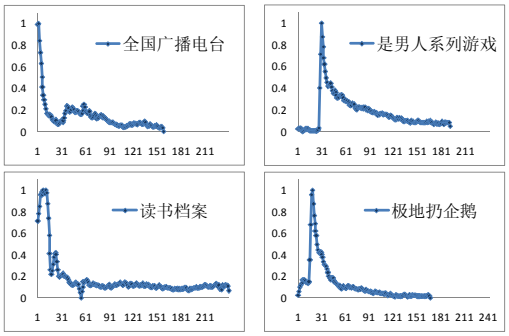


图 4 使用 FastDTW 度量不等长序列距离结果

3.4 改进的 K 中心点聚类

（1） K 中心点聚类

K 中心点聚类方法是基于划分的聚类方法的一种，给定数据样本和簇的数量 K，划分方法将聚类对象划分到 K 个簇，并以一个优化目标为准则，给定样本的距离或相似度度量，最终簇内样本距离较小，簇间样本距离较大。考虑现实数据中离群点的影响，本研究采用 K 中心点方法以规避极端值对数据聚类的扭曲。

K 中心点算法的目标是最小化优化目标绝对误差标准 $E=\sum_{k=1}^{K}\sum_{p\in C_k}|p-o_k|$，其中 $|p-o_k|$ 为空间中的点 p 与所属簇 C_k 代表对象 o_k 之间的距离。此算法反复迭代的最终结果是簇的代表对象成为簇的实际中心点或者最接近中心点的对象。

本文采用的典型 K 中心点聚类方法 PAM(Partioning Around Mediods)的基本算法步骤如下：

算法 1 K 中心点聚类

输入：

K：簇的个数，

D：包含 n 个对象的样本集合，

$Max_{iteration}$ ：最大迭代次数。

输出：K 个簇。

方法：

1 随机选择 D 中的 K 个样本作为初始簇代表对象；

2 将每个剩余样本分配到距离最近的簇；

3 随机选择一个非代表对象 O_{new}；

4 计算使用 O_{new} 代替簇代表对象的总代价（优化目标的变化） $S=\Delta E$；

5 (5) if $S<0$ and 未达到迭代次数 $Max_{iteration}$ ，then 使用 O_{new} 替换旧的簇代表对象，

6 else 重复 2 到 5

（2）簇中心计算

基于距离的聚类方法如 K 中心点、K 均值需要在每次更新时重新计算每个簇的中心，使用 FastDTW 算法尽管能解决不等长序列的距离度量问题，但只有簇中心被确认时才能计算样本与簇中心的距离。

传统的 K 中心点算法将每个划分到簇的样本特征进行平均，取得均值作为簇中心，以距离簇中心最近的点作为簇的代表样本。但对于本文中的不等长时间序列样本，采用直接求均值的方法显然不可行。因此，计算簇中心需要对数据进行变换。

将时间序列样本变换为等长序列的基本方法是对序列进行缩放，等间隔采样序列中的数据点或者进行插值以缩小或放大序列长度。但此种方法存在以下缺陷：

① 可能会使得序列中的一些重要数据点（如拐点）被删除；

② 缩放后的序列曲线与原始数据基本保持分布一致，在进行簇中心计算时，同类趋势但时域分布不同的曲线最终可能被平滑为平稳的数据分布。

为解决以上问题，我们借鉴数字图像处理中的 Seam Carving[6]方法并将其应用到一维数据上。Seam Carving 的基本思想是分别计算每个像素点的能量函数，函数值反映了该像素点的重要程度，依次迭代删除或插值能量值最低的像素点来实现图像的缩放并保证图像的主要信息被保留下来。

在一维时间序列上，假定长度为 L 的初始序列 $S=\{d_1,d_2,...,d_i,...,d_L\}$，

我们希望将其变换为新的长度为 L^{new} 的序列 S^{new}，

定义第 i 个数据点的能量函数 $E(i)=\sum_{j=i-r}^{j=i+r}\left|d_i-d_j\right|$，

其中，r 为考察半径，能量函数的物理意义即是该数据点邻域的变化强烈程度。

序列变换的算法如下：

算法 2 时间序列数据变换

输入：

序列 $S=\{d_1,d_2,...,d_i,...,d_L\}$，平滑窗口 W，

宽度 L_W

目标长度 L^{new}，半径 r

输出：目标序列 S^{new}

方法：

While S序列长度不等于L^{new} {

计算每个数据点的能量值 $E(i)=\sum_{j=i-r}^{j=i+r}\left|d_i-d_j\right|$，

取得能量值最小的数据点 t，

if(S 长度小于 L^{new}) 以 d_t 进行插值

else if(S 长度大于 L^{new}) 删除数据点 t

}

序列数据变换使得簇的中心计算变得可能，给定 K 个 Cluster，对于 $C_k, k = 1,2,....,K$，有序列集合 $Q_k = \{S_{k1}, S_{k2}, ..., S_{km}, ..., S_{kM}\}$ 为被聚集到该簇的样本序列，则新的簇中心可计算为

$$Centriod_k = \frac{1}{M}\sum_{m=1}^{M} S_{km} \quad (4)$$

4 试验和结果分析

我们使用前述 K 中心点方法将 SNS 应用聚集到 k=7 个类别。依据聚类结果的时序统计数据曲线形状，我们可以将数据集中的 SNS 应用划分为以下几种类型并利用产品生命周期理论[5]进行分析。

产品的时序统计数据根据不同的发展阶段，可以以生命周期的形式表现出来。以 SNS 开放平台应用的活跃用户数量为例，产品生命周期理论可以曲线划分为 4 个阶段：

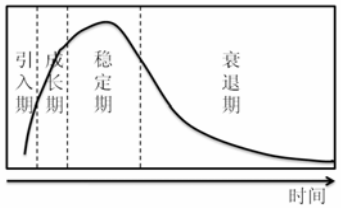


图 5 产品生命周期曲线

① 引入期：应用程序初步进入市场；
② 成长期：应用程序在 SNS 平台上处于累积用户群的阶段，也是在线应用成熟和被市场选择的时期。
③ 稳定期：处于稳定期的 SNS 开放平台应用程序在用户群的规模上达到了峰值，也是商业价值最大化的时期。
④ 衰退期：处于衰退期的应用程序逐渐流失 SNS 用户，用户黏着度降低，成为不活跃的应用程序至被市场淘汰。

根据对观察窗口（140-250 天）内的 SNS 开放平台应用统计数据的聚类结果在产品生命周期中的分布，此类在线应用程序统计曲线有以下几种典型类型：

① 快速衰退型。此类的 SNS 开放平台应用程序在观察时间区间内基本处于产品的衰退期，并且具有较快的衰退速度。
② 快速增长-衰退型。应用程序有一个较短的成长期和快速衰退的产品生命阶段。这种类型的产品包括短期热门的或者缺少后续维护研发的应用程序等。

③ 平稳增长型。应用程序在观察时间区间内处于引入期、成长期和平稳期前期。

④ 平稳增长-衰退型。在观察时间区间内应用程序至少具有平稳的增长期、平稳期和衰退期。平稳增长-衰退型的 SNS 开放平台应用通常是具有一定商业或应用价值的软件。

表 2 聚类结果中的 SNS 平台应用类型

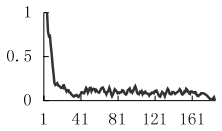
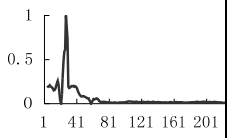
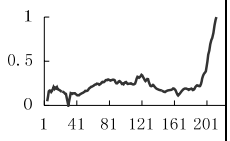
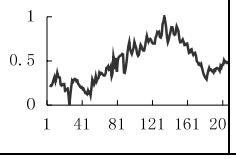
类型	代表曲线	样本数量
快速衰退型		272
快速增长-衰退型		186
平稳增长型		20
平稳增长-衰退型		63

表 2 给出了聚类分析得出的 4 种 SNS 开放平台活跃用户量的分布类型，其中那个曲线的横轴为时间，数值轴的数量被规范化到[0,1]区间。可以看出，快速衰退型或者快速增长-衰退型的应用数量占据了较大比例，反映出基于 SNS 开放平台的应用程序仍然在需求调研、可行性分析、系统实施维护方面存在缺陷，只能在短期内吸引用户，而不能产生较高的用户黏着度。相比之下，能够在长时间内保持平稳增长或具有较平稳的生命周期的应用产品数量则非常少。这一方面是由 SNS 平台的用户流动性特征决定，另一方面也显示了成熟的应用的缺失。

5 结论及下一步工作

SNS 开放平台应用在软件与用户群组中形成的特殊现象尚缺少理论与技术探讨。本文首先对 SNS 开放平台及其应用的相关概念进行了介绍，并总结和分析了 SNS 开放平台和应用的现状及运营模式。基于已有数据，我们使用数据挖掘的方法，提出使用 DTW 算法度量应用软件统计数据序列距离，并依此进行聚类分析来挖掘开放平台应用的生存模式和生命周期。

本文实验所分析的对象仅仅是国内主流开放平台的应用，虽然对数据进行了筛选，但要获取更可靠的结果，仍需对国内外主要 SNS 开放平台的数据进行跟踪分析。此外，还可从软件创新理论、传播学理论等角度尝试进一步验证和研究。

参考文献

[1] 史训纲.多媒体挖掘讲义：Dynamic Time Warping. [2009-02-10]. http://dblab.cs.nccu.edu.tw.

[2] Han J W, Kamber M. Data Mining: Concepts and Techniques. San Francisco: Morgan Kaufmann Publishers, Inc, 2001.

[3] Salvador S, Chan P. FastDTW: Toward Accurate Dynamic Time Warping in Linear Time and Space, Intelligent Data Analysis, 2007, 11(5): 561-580.

[4] Sankoff D, Kruskal J. Time Warps, String Edits, and Macromolecules: the Theory and Practice of Sequence Comparison. Addison-Wesley Pub. Co., Advanced Book Program, 1983

[5] Chang P T, Chang S H. A Stage Characteristic-Preserving Product Life Cycle Modeling. Mathematical and Computer Modeling, 2003, 37(12-13): 1259－1269.

[6] Avidan S, Shamir A. Seam Carving for Content-Aware Image Resizing. ACM SIGGRAPH 2007 papers. San Diego, California: ACM, 2007, 26(3): 10.

[作者简介]

裴珊珊，女，1984 年生，硕士研究生，发表论文 1 篇。

叶小梁，女，1952 年生，研究员，硕士生导师，发表论文 60 余篇。