文章编号:1671-9352(2014)09-0150-04 DOI:10.6040/j.issn.1671-9352.2.2014.235

基于 KNN 的 Android 智能手机微信取证方法

吴熙曦[1,2],李炳龙[1,2],张天琪[3]
(1. 信息工程大学四院,河南 郑州 450004;
2. 信息工程大学数学工程与先进计算国家重点实验室,河南 郑州 450004;
3. 中国人民解放军65547部队,辽宁 鞍山 114200)

摘要:针对微信数据多,无法从中快速找到与案件相关数据的问题,提出了一种基于KNN(k-nearest neighbor)算法的Android智能手机微信取证方法。引入词语相似度计算会话间的距离,将微信会话表示成特征词的向量,用KNN算法对会话进行分类,迅速找到与犯罪有关的聊天内容,并通过实验验证了该方法的可行性与准确性。

关键词:微信取证;数据挖掘;KNN算法;词语相似度
中图分类号:TP311 **文献标志码**:A

A KNN based forensic method of Android smartphone WeChat

WU Xi-xi[1,2], LI Bing-long[1,2], ZHANG Tian-qi[3]
(1. The Forth College, Information Engineering University, Zhengzhou 450004, Henan, China;
2. State Key Laboratory of Mathematics Engineering and Advanced Computing, Information Engineering University, Zhengzhou 450004, Henan, China; 3. No. 65547 Unit, PLA, Anshan 114200, Liaoning, China)

Abstract: To solve the problem that data of WeChat is so much that data related to the case can't be found quickly, a Android smart phone WeChat forensic method based KNN algorithm was presented. Word similarity was introduced to calculate the distance of conversations. The conversations would be represented as a vector of feature words and categorized with KNN algorithm to quickly find the crime-related data. The experiments verify the feasibility and accuracy of the method.

Key words: WeChat forensics; data categorization; KNN algorithm; word similarity

0 引言

随着移动互联网的快速发展,智能手机已经成为人们生活中不可缺少的一部分。与此同时,利用智能手机实施的犯罪活动也屡见不鲜。传统的手机取证通常是对手机中的短信、通话记录等信息进行取证。随着微信等社交软件的兴起,利用社交软件进行的犯罪活动也变得越来越猖獗。微信是腾讯公司于2011年1月21日推出的一个为智能手机提供即时通讯服务的免费应用程序,微信支持跨通信运营商、跨操作系统平台通过网络快速发送免费(需消耗少量网络流量)语音短信、视频、图片和文字,同时,也可以使用通过共享流媒体内容的资料和基于位置的社交插件"摇一摇"、"飘流瓶"、"朋友圈"、"公众平台"等插件服务。据腾讯内部消息,截至2013年10月24日,微信的使用人数已超过6亿。微信自己也宣称:微信,是一种生活方式。正是由于微信的流行,对微信的取证显得尤为重要。

人们在使用微信进行聊天、交友、上传照片时产生了大量的数据,取证人员要对这些大量繁杂的数据进行分析,从中找出与案件相关的数据,需要一种

收稿日期:2014-06-24; **网络出版时间**:2014-08-28 17:00
网络出版地址:http://www.cnki.net/kcms/doi/10.6040/j.issn.1671.9352.2.2014.235.html
基金项目:国家自然科学基金资助项目(60903220);郑州市科技攻关项目;信息工程大学硕士论文基金资助项目
作者简介:吴熙曦(1990-),女,硕士研究生,研究方向为智能手机取证. E-mail:1107662015@qq.com

行之有效的方法。KNN((k-nearest neighbor))算法[1]由 Cover 和 Hart 于 1968 年提出,是一种基于实例学习、非参数的分类技术,简单易行,分类效果良好,对不同数据集都有很好的可操作性,被广泛应用于基于统计的机器学习中。文献[1]提出了一种改进的 KNN 算法,引入 kdtree 存储结构并通过样本聚类方法来减小样本空间,以此来提高 KNN 算法在文档分类时的运行时间。同样地,文献[2]根据 SOM(自组织映射神经网络)理论、特征选取和模式聚合理论,提出一种改进的 KNN 文本分类方法。应用特征选取和模式聚合理论有效地降低了特征空间维数,提高了文本分类的精度和速度。文献[3]提出一个简单的非特征加权 KNN 文本分类算法,使用特征选择方法,找出相关的特征使用的特征交互方面(基于词的相互依存关系),能够大大减少特征数。可见 KNN 算法在文本分类中的重要性。

Android 系统作为统领智能手机市场的一个移动操作系统,拥有巨大的用户数量。本文提出了一种基于 KNN 的 Android 智能手机微信取证方法,利用聚类算法对微信聊天内容进行分类,从大量的聊天数据中挖掘出与案件相关的数据,引入词语相似度,计算会话之间的距离,为取证人员提供了一种有效的微信聊天分类方法。

1　聚类算法

KNN 算法是聚类算法中的一种重要算法,聚类是指将物理或抽象的集合分组成为由类似的对象组成的多个类的过程[4]。由聚类生成一组数据对象的集合,这些对象与同一个簇中的对象彼此相似,与其他簇中对象相异。在许多应用中,可以将一个簇中的数据作为一个整体来看待。聚类是在事先不规定分组规则的情况下,将数据按照其自身特征划分成不同的群组。要求在不同群组的数据之间有明显差别,而每个群组内部的数据之间尽量相似。

KNN 算法的思想是:计算一个点 A 与其他所有点之间的距离,取出与该点最近的 k 个点,然后统计这 k 个点里面大部分的点所属的分类,则点 A 属于该分类[1]。

假设有 c 个类:$\omega_1,\omega_2,\cdots,\omega_C,k_1,k_2,\cdots,k_c$ 分别是类 $\omega_1,\omega_2,\cdots,\omega_C$ 的样本个数,则判定函数可定义为

$$\mu_i(A)=k_i,i=1,2,\cdots,c。\tag{1}$$

根据公式(1),判定规则为:如果 $\mu_j(A)=\max k_i$,则 $A\in\omega_j$。

假设 A 包含 m 个样本,每个样本有 n 个分量:$X_i=(x_{i1},x_{i2},x_{i3}\cdots x_{in}),i=1,2,3\cdots m$。样本 $Y=(y_1,y_2,y_3\cdots y_n)$ 是待分类的样本,计算 Y 与每个样本的欧几里得距离:

$$D(Y,X_i)=\sqrt{\sum_{j=1}^{n}(y_j-x_{ij})^2}。\tag{2}$$

选出 k 个距离 Y 最小的样本,则这 k 个样本就是与 Y 最邻近的样本,统计出这 k 个样本所属的分类,则 Y 属于样本数量最多的分类。

2　词语相似度计算

2.1　会话的向量空间表示

为了对微信的聊天内容进行聚类分析,将微信聊天记录转化成向量的形式。例如与某人的会话表示为 $\boldsymbol{C}:(W_1,W_2,\cdots,W_n)$,其中 W_i 是第 i 个特征项的权重;n 表示会话 $\boldsymbol{C}$ 中特征项的个数。特征项是指出现在聊天中并能表示该会话特点的基本语言单位。在聊天中一般选择文字中的字、词或词组作为特征。特征项的权重 W_i 是指能够代表会话 $\boldsymbol{C}$ 能力的大小,它体现了该特征在会话 $\boldsymbol{C}$ 中的重要程度。要将会话表示为一个向量,首先要将会话分词,由这些词作为向量的维数来表示聊天内容。最初的向量表示使用布尔权重的计算方法,即会话中出现了该词,则向量中该维为 1,否则为 0。但这种方法无法体现这个词在会话中的作用程度,所以本文采用项频率逆文档频率公式[2]来计算:

$$w(t,c)=\frac{\mathrm{tf}(t,c)*\log\left(\frac{N_c}{\mathrm{cf}}\right)}{\sqrt{\sum_{i=1}^{N_t}\left[\mathrm{tf}_i(t,c)*\log\left(\frac{N_c}{\mathrm{cf}_i}\right)\right]^2}}。\tag{3}$$

其中,$w(t,c)$ 表示特征词 t 在会话 c 中的权重,N_c 表示会话总数,N_t 表示特征词总数,$\mathrm{tf}(t,c)$ 表示特征词 t 在会话 c 中的词频,cf 表示训练会话集中出现词 t 的会话数。

2.2　基于同义词词林的词语相似度

词语的相似度[5]与词语的词法、句法、语义等许多特点有关,影响相似度的两个重要指标是词语相似性和词语相关性。一般用词语间的语义距离衡量词语的相似性。《同义词词林》是梅家驹等人于 1983 年编纂而成,这本词典中不仅包括了一个词语的同义词,也包含了一定数量的同类词,即广义的相关。《同义词词林》词典分类采用 5 层级体系,具备 5 层结构,如图 1 所示。

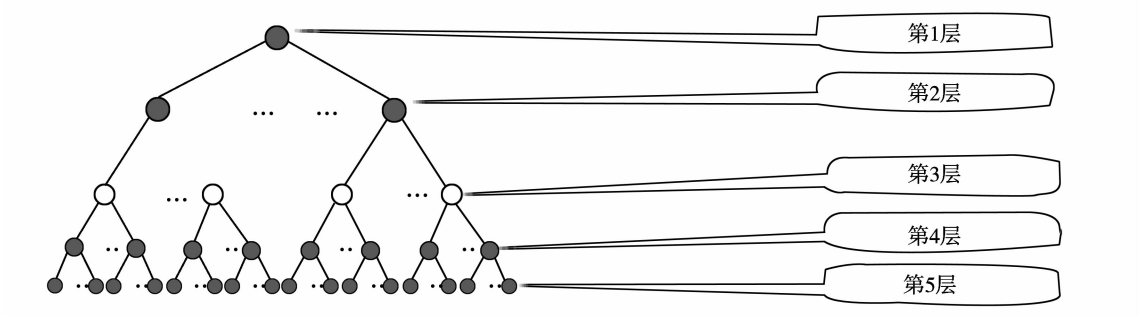

图1 同义词词林5层结构图
Fig. 1 Five-layer structure of Synonyms

随着级别的递增，词义刻画越来越细，到了第5层，每个分类里词语数量已经不大，很多只有一个词语，已经不可再分，可以称为原子词群、原子类或原子节点。

基于同义词词林的义项相似度的主要思想是：基于同义词词林结构，利用词语中义项的编号，根据两个义项的语义距离，计算出义项相似度。两个义项 A 和 B 的相似度[6]用 Sim 表示：

$$\mathrm{Sim}(A,B)=a*\cos\left(n*\frac{\pi}{180}\right)\left(\frac{n-k+1}{n}\right),\quad(4)$$

其中，a 是系数，n 是分支层的节点总数，k 是两个分支间的距离。

计算词语相似度时，把两个词语的义项分别两两计算，取最大值作为两个词语的相似度值。例如：词语“骄傲”的义项编号有“Da13A01 =”、“Ee34D01 =”；“仔细”的义项编号有“Ee26A01 =”、“Ee28A01 =”。分别计算义项的相似度为0.1，0.1，0.483 920，0.510 077。可以得出“骄傲”和“仔细”的词语相似度为0.510 077，即4个相似度的最大值。

3 基于 KNN 的微信会话分类

3.1 会话间相似度计算

对微信会话进行分类之前，应先计算会话间的相似度。首先对训练集中的所有会话进行特征选择，选取出特征词，利用公式(3)统计出特征词的权重。然后利用基于同义词词林的词语相似度计算得出特征词间的相似度，设相应的阈值为0.8，即相似度大于0.8的词语视为同义词，权重较大者为代表特征词，权重较小者取其权重与相似度之积为其权重，最后用向量表示会话。

以下是从两个有主题的会话中提取的特征词及权重，权重由公式(3)计算得：

(1) 房产(0.6)，贷款(0.4)，投资(0.3)，风险(0.3)；

(2) 信用卡(0.5)，房子(0.5)，抵压(0.3)，危险(0.2)。

这8个特征词分别记 $w_1,w_2,w_3,\cdots,w_8$。经计算两个词之间的相似度大于阈值的有：

$\mathrm{Sim}(w_1,w_6)=1.00$，$\mathrm{Sim}(w_4,w_8)=0.95$。

即“房子”和“房产”相似度为1，两者任取一个均可；而“风险”与“危险”相似度为0.95，取权重较大者“风险”，则会话(1)中“风险”权重仍为0.3，会话(2)中“危险”的权重则为0.95与0.2之积0.19。

3.2 算法流程

图2是基于 KNN 的微信分类算法流程。

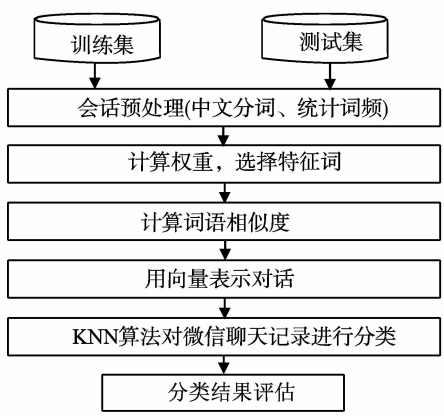

图2 会话分类流程
Fig. 2 Process of session classification

(1) 对聊天会话进行预处理：中文分词，统计词频。预处理的任务是去掉文本中与分类无关的部分。对于中文文本而言，词与词之间并没有十分明显的切分标志，所以需要进行分词。目前的分词方法主要分为基于理解的方法、基于字符串匹配的方法以及基于统计的方法。在分词时还要对其进行词性的标注、人名以及地名等这些专有名词的识别。因为计算机并不具备人类的智能，为了使计算机能够进行后续的处理，微信的聊天记录需要一种较为有效的表示方式。

(2) 用公式(3)计算出所有词的权重，根据权重值选择特征词。目前计算机能够识别的对象为0和1，因此聊天内容在被计算机处理之前，需要将聊天

转换成 0 和 1 代码的形式。在一个会话中往往包含很多特征项,这些特征项中有一些对会话分类有价值,有一些对会话分类没有价值,如何才能辨别某一特征项是否对会话分类有价值,在这里提出了权重这一概念,通过计算权重的大小来区分有价值或者没有价值的特征项。

(3)计算词语相似度。基于同义词词林计算各类特征词的相似度,选择最优特征词。

(4)用特征词的权重向量来表示会话。

(5)用 KNN 分类算法来实现会话的分类,即分为正常聊天或可疑聊天(与犯罪有关)。

(6)用召回率、准确率、查全率等方法对分类结果进行评估。会话分类系统评估的标志是分类器对会话分类的速度与分类的准确度。分类器的分类速度取决于分类器的计算公式是否简约。分类的准确度在于使用该分类器对会话分类的结果与人工分类结果相比较,当两者结果相似度越高,则说明分类器的准确性就越高。评估分类器 3 个指标分别为运算速率、准确率以及召回率。

准确率计算公式:

$$准确率(p)=\frac{分类正确的会话数}{实际分类的会话数}。\qquad(5)$$

召回率计算公式:

$$召回率(R)=\frac{分类的正确会话数}{应有会话数}。\qquad(6)$$

准确率与召回率是一种此消彼长的关系,即分类要想获得较高的准确率,往往就要牺牲召回率;同样,就要牺牲准确率。因此,$F1$ 值就是一种使得文本分类准确率与召回率达到平衡的计算方法,进而对分类结果进行测评。$F1$ 值计算公式为

$$F1=\frac{准确率*查全率*2}{准确率+查全率}。\qquad(7)$$

4　实验及结果

实验数据取自 5 部实验所用 Android 智能手机中与 800 个微信好友的聊天记录,这 800 个会话内容分别包括祝福语、广告、体育等 7 个类别,正常聊天和与犯罪相关的都有。另外有 800 组类别相同且数量相同的微信对话作为训练集。

首先对聊天会话进行中文分词,过滤主题无关词,计算词频。选择词频最高的 156 个词作为特征词,建立会话向量。KNN 算法中 k 的取值很重要,如果 k 值太小,则 KNN 分类器容易受到优于训练数据中的噪声而产生过分拟合的影响;相反,如果 k 太大,KNN 分类器可能会误分分类测试样例,因为最近邻列表中可能包含远离其近邻的数据点。经过实验发现 $k=15$ 时分类效果最好,表 1 是实验数据。

表 1　实验数据

Table 1　Experimental data

类别	训练集	测试集
祝福	80	80
广告	130	130
体育	100	100
财经	120	120
笑话	90	90
科技	150	150
犯罪相关	110	110

准确率、查全率、$F1$ 都是文本分类的评价方法,从表 2 可以看出,这三种结果都比较高,说明 KNN 算法应用于微信聊天内容分类是可行的并且准确率较高。但是测量结果总会有一定误差,这是因为当样本不平衡时,有可能导致当输入一个新样本时,该样本的 K 个邻居中大容量类的样本占多数。因为算法只计算“最近的”邻居样本,若某一类的样本数量很大,那么或者这类样本并不接近目标样本,或者这类样本很靠近目标样本,从而导致分类结果不准确。

表 2　分类结果评估

Table 2　Assessment of classification results

结果	准确率(P)	查全率	$F1$
宏平均	0.937	0.908	0.922
微平均	0.924	0.938	0.931

5　结论

本文利用同义词词林,对微信聊天内容进行词语相似度计算,选出最佳特征词,然后用向量表示会话,再用 KNN 算法对会话进行分类,能够较快地从大量聊天记录中找到与案件相关的数据。但从上述实验结果可以看出 KNN 算法也存在缺陷,因此下一步的研究工作主要是针对 KNN 算法在微信会话分类时存在的不足,提出有效的改进办法。

参考文献:

[1] JIANG Zongli, YI Deng. Improving KNN based text classifications[C]//Proceedings of the 2nd International Conference on Future Computer and Communication (ICFCC 2010). Piscataway: IEEE, 2010:317-337.

[2] 钱晓东,王正欧. 基于改进 KNN 的文本分类方法[J]. 情报科学,2005, 23(4):550-554.
QIAN Xiaodong, WANG Zhengou. Text classification method based on improved KNN[J]. Information Science, 2005, 23(4): 550-554.

(下转第 165 页)

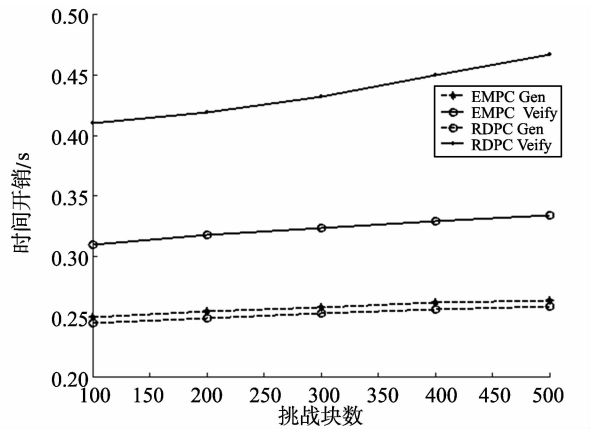

图 3　EMPC 和 RDPC[7] 时间开销比较(单副本挑战)

Fig. 3　The compare between EMPC and RDPCfor time overhead

5　结束语

本文提出了一种利用同态哈希技术的多副本持有性证明方案(EMPC),能够同时对多个副本的持有性进行验证。通过改进 Merkel 哈希树,使其支持动态数据操作,且利用 γ 编码技术减少数据块的验证和更新等操作中带宽消耗。安全性分析证明了其具有抵抗替换、重放和伪造攻击的能力。并与文献[7]中的 RDPC 进行比较,结果表明本文的方案无论是在安全性、通信开销还是时间开销方面都是更优的。

参考文献:

[1] JUELS A, KALISKI JR B S. PORs: proofs of retrievability for large files[C]//Proceedings of the 14th ACM Conference on Computer and Communications Security. New York: ACM Press, 2007: 584-597.

[2] ATENIESE G, BURNS R, CURTMOLA R, et al. Provable data possession at untrusted stores[C]//Proceedings of the 14th ACM Conference on Computer and Communications security. New York: ACM Press, 2007:598-609.

[3] ERWAY C, KÜPCÜA, PAPAMANTHOU C, et al. Dynamic provable data possession[C]//Proceedings of the 16th ACM Conference on Computer and Communications Security. New York: ACM Press, 2009: 213-222.

[4] CURTMOLA R, KHAN O, BURNS R, et al. MR-PDP: multiple-replica provable data possession[C]//Proceedings of the 28th International Conference on Distributed Computing Systems (ICDCS'08). Los Alamitos: IEEE Computer Society, 2008: 411-420.

[5] CHEN L X. A homomorphic hashing based provable data possession[J]. Journal of Electronics and Information Technology, 2011, 33(9): 2199-2204.

[6] 李超零,陈越,谭鹏许,等. 基于同态 Hash 的数据多副本持有性证明方案[J]. 计算机应用研究,2013,30(1):265-269.
LI Chaoling, CHEN Yue, TAN Pengxu, et al. Multiple-replica provable data possession based on homomorphic hash[J]. Application Research of Computers, 2013, 30(1):265-269.

[7] CHEN Lanxiang, ZHOU Shuming, HUANG Xinyi, et al. Data dynamics for remote data possession checking in cloud storage[J]. Computers & Electrical Engineering, 2013, 39(7): 2413-2424.

[8] KROHN M N, FREEDMAN M J, MAZIERES D. On-the-fly verification of rateless erasure codes for efficient content distribution[C]//IEEE Symposium on Security and Privacy. Los Alamitos: IEEE Computer Society, 2004:226-240.

[9] WANG Qian, WANG Cong, REN Kui, et al. Enabling public auditability and data dynamics for storage security in cloud computing[J]. IEEE Transactions on Parallel DistribSyst, 2011, 22(5):847-859.

[10] ATENIESE G, DI PIETRO R, MANCINI L V, et al. Scalable and efficient provable data possession[C]//Proceedings of the 4th International Conference on Security and Privacy in Communication Netowrks. New York: ACM Press, 2008: 1-11.

(编辑:许力琴)

(上接第 153 页)

[3] SOUCY P, MINEAU G W. A simple KNN algorithm for text categorization[C]//Proceedings of IEEE International Conference on Data Mining (CDM 2001). Washington: IEEE Computer Society, 2001: 647-648.

[4] 杨莉莉. 基于数据挖掘的数字取证模型设计[J]. 南京师范大学学报,2006, 29(6):18-21.
YANG Lili. Design of digital forensics model based on data mining[J]. Journal of Nanjing Normal University, 2006, 29(6):18-21.

[5] 鲁婷,王浩,姚宏亮. 一种基于中心文档的 KNN 中文文本分类算法[J]. 计算机工程与应用,2011, 47(2):127-130.
LU Ting, WANG Hao, YAO Hongliang. A KNN Chinese text classification algorithm based on center document [J]. Computer Engineering and Applications, 2011, 47(2):127-130.

[6] 田久乐,赵蔚. 基于同义词词林的词语相似度计算方法[J]. 吉林大学学报:信息科学版,2010, 28(6):602-608.
TIAN Jiule, ZHAO Wei. Words similarity algorithm based on tongyici cilin in semantic web adaptive learning system[J]. Journal of Jilin University: Information Science Edition, 2010, 28(6):602-608.

(编辑:许力琴)