DOI:10.11721/cqnuj20170454

基于词嵌入的云存储可搜索加密方案*

钟 晗[1], 郭 飞[2]
(1. 南京大学 计算机科学与技术系, 南京 210064; 2. 重庆大学 计算机学院, 重庆 400033)

摘要:【目的】提出基于词嵌入的云存储可搜索加密方案,视图解决云储存的加密数据的管理,并高效地检索加密数据这一难题。该方案的优势在于克服了加密方案不能进行语义搜索的难题。【方法】通过建立高维关键词的词嵌入,增加语义距离扩展关键词集的方式建立安全索引,并用伪随机函数对私钥和关键词进行安全保护。【结果】此设计方案既保证了数据存储的安全性,又提供了数据检索的灵活性,安全检索模型避免了用户检索过程中关键信息的泄露。【结论】采用全同态加密验证了方案的有效性。在维基百科数据集上进行测试表明维度越高的词嵌入搜索精确率越高,同时开销也随之增大。

关键词:云存储;可搜索加密;词嵌入;全同态加密
中图分类号:TP309 **文献标志码**:A **文章编号**:1672-6693(2017)04-0070-05

1 云存储的优缺点

云存储的应用主要满足于企业及个人用户远程数据存储与管理,以应对移动互联与大数据的挑战。云存储依附于云计算,云计算的高性能计算服务能为诸如文档检索、文档分析方面提供支撑。对于云存储服务提供商来说,在云端存储管理中整合所有存储资源,能有效避免资源浪费。对于用户来说,把本地数据迁移到云端存储,不仅减少了数据维护,节省了开销,而且可以随时实现数据的共享和发布。

用户将数据存储于云端,数据则脱离了用户对物理介质的管控,因此数据存储的安全性显得至关重要。用户总是希望使用更安全稳定的云储存服务来保证数据安全和用户隐私。然而,诸如管理员的不当操作或者黑客等非法手段访问获取数据导致数据泄露的事故时有发生[1-3]。因此,应建立完备的数据存储安全保护机制。目前,云存储有 4 类存储安全服务:认证服务、数据加密存储、安全管理和安全日志与审计,其中存储加密服务是最关键的业务。

为了解决数据加密存储问题,数据应采取密文存储的方式存于云端。数据加密又分为用户加密和云端加密。用户加密的优点在于云端数据即使被泄露也能保证数据不会被轻易破解,安全性较高;云端加密的优点在于云存储提供的加密方式对于用户来说是透明的,用户不必考虑加密的细节问题。云端加密的缺点在于难以防范内部数据的泄露。因此,对于重要的文档数据,常用的方法是由用户将数据进行加密处理,再把密文存储于云端存储中[4];但对于云端加密数据的各类操作不能像明文一样简单。如何从大量加密文档中搜索符合条件的文档变得相当棘手。如果用户下载大量加密数据再将它们解压后搜索,势必会浪费大量资源;最理想的是直接在云端搜索返回想要的结果。因此可搜索加密机制成为一种可行的云端密文检索技术,它在保护数据安全性的同时能有效降低加密、解密等服务计算的开销。

2 可搜索加密机制

可搜索加密的工作原理为:当用户需要搜索加密文档的信息时,只需把关键词或搜索表达式处理为搜索凭证,将搜索凭证与加密后的文档或建立的密文文档索引进行算法匹配;如果匹配成功,则返回相应文档,再将文档解密即可。在整个搜索过程中执行方并不能获取检索内容和明文内容,既保证了检索的可操作性又保证了文档的安全性[5-6]。

目前可搜索加密的加密算法分为 2 大类,分别基于对称密码学原理与基于公钥密码学原理。前者由随机生成函数、哈希算法和对称算法构造;它的消息发送方和消息接收方必须使用相同的密钥,如 DES,AES。而基于

* **收稿日期**:2017-02-07 **修回日期**:2017-05-08 **网络出版时间**:2017-06-15 11:23
资助项目:教育部人文社科项目(No.16JDSZ2019);重庆市教委科学技术研究项目(No.KJ1500918)
第一作者简介:钟晗,男,研究方向为信息安全,E-mail: 648505893@qq.com;**通信作者**:郭飞,编辑,E-mail:fliggy@163.com
网络出版地址:http://kns.cnki.net/kems/detail/50.1165.N.20170615.1123.010.html

公钥密码学原理的加密算法即非对称加密，则使用 2 个不同的密钥：1 个公共密钥 PK 和 1 个私有密钥 SK；它是利用双线性映射数学工具构建而成的加密方式，如 RSA，Elgamal 等。Song 等人[7]提出基于对称密码算法的分块加密：当需要搜索关键词时，将最终密文与关键词进行异或运算，如果匹配则成功返回。但是它的算法效率不佳，解密算法与密文大小呈线性关系。针对此问题，Goh 在文献[7]基础上利用独立哈希函数与 BF 数据结构建立安全索引的方法实现海量快速检索[8]。Boneh 等人[9]提出非对称密码学的加密搜索算法，即以非对称加密方式对明文关键词进行公钥加密，然后生成可用公钥检索的密文信息。之后 Curtmola 等人[10]为了提高基于公钥算法搜索的安全性，提出限制只具有私钥的用户才能用公钥加密的数据进行搜索。文献[11]通过混合云环境中的多用户数据共享尝试解决此类问题，保证数据的完整性。可搜索加密发展至今，它的核心问题就是搜索凭证与加密后的文档通过某种计算仍能得到相应结果。

3 问题描述

3.1 系统结构

云环境下的加密存储和检索主要由两大部分构成，即客户端和云存储端，图 1 表明了它们之间的交互关系。

1）客户端即用户端。它拥有明文的访问权与读写权，能通过密钥对明文及明文索引进行加密和密文的解密工作，负责与云存储端的交互，能上传密文和密文索引数据，能下载密文数据，能发送检索、更新、删除等请求。

2）云存储端即服务器端。它提供认证服务、安全管理、日志审查、密文存储、密文索引管理等功能。云存储系统的结构模型由 4 层组成，从下到上分别是存储层、基础管理层、应用接口层和访问层。存储层主要负责密文、密文索引的存储和空间的分配，数据的备份与恢复工作。该层对于上面 3 层与外界都是透明的，它可以是存储集群，也可以是直连式存储。基础管理层主要负责存储层中密文索引管理工作，同时还负责密文及密文的数据优化压缩和策略控制，具体任务包括密文冗余数据的合并和密文索引的创建、更新、删除。应用接口层负责加密搜索算法的创建和计算。该层的应用接口可根据用户的需求，生成相应的密文计算算法。访问层提供认证服务，主要负责与客服端的交互和认证，是对外唯一的访问入口。分层设计的好处之一是由于各层独立，每一层只能与上下两层交互，保证了数据和操作的隔离，确保了数据的安全性；二是如果某一层需要修改或替换，对整个云存储结构不会更改且易于扩展。

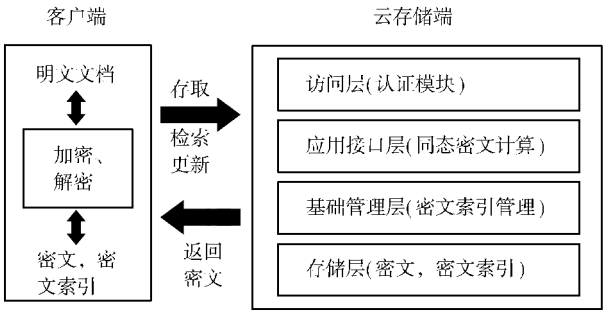

图 1　云存储加密存储及检索方案

Fig. 1　Cloud storage encrypted storage and retrieval design

加密存取及检索过程如下。

1）存取过程。存取过程分为存储密文数据（存储过程）和密文索引、读取密文数据（读取过程）两类操作。存储过程为：客户端负责把明文文档进行加密，同时文档中的词干和词频一并建立索引加密为密文索引提交到云存储端。云存储端在收到客户端的交互请求后，把密文文档和密文索引文档存储于存储层，并在基础管理层更新索引，管理密文索引。读取过程为：客户端在完成与云存储端的认证交互请求后，向云存储发送需要的密文索引。首先由基础管理层查询是否有符合请求的密文索引，然后层层提交，并交付客户端，客户端收到密文后，解密得到明文信息。

2）检索、更新、删除过程。检索过程为：客户端在接受用户需要检索的关键词或查询条件的请求后，加密生成相应的搜索凭证；并在访问层授权认证后把授权凭证提交至应用接口层。应用接口层就搜索凭证与密文索引进行计算，根据密文索引的计算结果找到对应的密文，以密文排序的方式返回客户端。客户端在收到密文后进行解密，即可得到相应的明文。对于云储存端，此类密文计算只是利用了云计算强大的计算能力。在计算过程中，云端并不能获取任何有用的明文或者索引信息，充分保证了数据的私密性。更新与删除操作相对简单一些。客户端提交更新密文和密文索引以覆盖云储存端对应的旧文件即可；删除过程为：根据密文索引删除对应的密文及密文索引即可。

3.2 威胁模型分析

此模型考虑潜在威胁对象有云端服务器和外部攻击者。云端服务器虽然能根据安全策略正确执行协议规范，但是作为第三方，服务器可能“好奇地”收集用户的搜索凭证，并根据校验凭证与返回索引结果来统计分析以获取额外的信息。而外部攻击者，可以伪装成授权的客户端对云存储进行查询攻击。因此，不能将所有的信息和操作交给云端，在设计搜索凭证方案时将遵循加密中所涉及的安全定义。

4 方案设计

4.1 词嵌入概述

词嵌入也称为词向量[12],它是利用深度学习大规模文本数据对每一个词建立高维的实数向量来表征词的含义。相对于基于传统词 TF-IDF 词频向量或 LSA 潜在语义向量,它能包含更多的词与词之间的含义。它的优点在于词嵌入可以通过计算2个单词的余弦距离来计算相似度,且语义相似度优于 TF-IDF 和 LSA。另外词嵌入完全是用户自学习的高维向量表示,与基于词典的加密搜索方案[13]相比,在没有获得单词与向量的映射之前是不能简单地推断出单词的。在统一的框架下,多语词嵌入可以进行跨语言检索。本研究采用 Word2vec[14]学习的词嵌入来建立搜索凭证,并完成语义检索。

4.2 符号定义

W:通过大规模语料库学习提取的单词嵌入集合,即字典 $W=(w_1,w_2,\cdots,w_n)$。

F:客户端所拥有的明文文档集合 $F=(F_1,F_2,\cdots,F_m)$。

C:通过加密后的密文集合,需上传到云储存端 $C=(C_1,C_2,\cdots,C_m)$。

T_w:搜索陷门,即用户输入搜索关键词并映射成词嵌入 w,单向函数生成的搜索凭证。

FID_w:明文文档集合 F 中包含关键词 w 的文件 ID 集合,对应的密文 ID 集合与 FID_w 对应的加密解密映射为 $FID_w \leftrightarrow CID_w$。

$I(T,ID)$:为保护用户隐私的语义关键词搜索而建立的索引函数,T 为搜索陷门,ID 为索引。

S_{T_w}:在云存储端,根据搜索陷门 T_w 进的搜索操作 S_{T_w},返回 CID 集合。

$f(x_{\text{key}},\cdot)$,$g(x_{\text{key}},\cdot)$:伪随机函数,定义为:$\{0,1\}^{*}\times x_{\text{key}}\rightarrow\{0,1\}^{l}$。

$Enc(x_{\text{key}},\cdot)$,$Dec(x_{\text{key}},\cdot)$:基于安全的加密/解密函数。

语义距离 $d(w_1,w_2)$,语义距离是单词相似度的一种语义描述。对于两个词嵌入 w_1 和 w_2,语义距离表示 w_1 和 w_2 相似程度,按照余弦夹角定义距离,$d(w_1,w_2)$ 越接近1表示语义越相近。$d(w_1,w_2)=\dfrac{w_1\cdot w_2}{\|w_1\|\cdot\|w_2\|}$,其中 $d(w_1,w_2)$ 的取值范围是$[-1,1]$。

$SP_{w,d}$:表示给定词嵌入 w 和 d,与词嵌入 w 的语义距离$\leqslant d$ 的其他词嵌入集合,即 $d(w,w'_i)\leqslant d$,$i\in\{1,2,\cdots,r\}$,$SP_{w,d}=\{w'_1,w'_2,\cdots,w'_r\}$。

$SP_{w,d}*$:$SP_{w,d}*$ 为满足词嵌入 w 语义距离$\leqslant d$ 的所有词嵌入集合,$SP_{w,d}*=w\cup SP_{w,d}$。

4.3 搜索方案

搜索请求:当客户端输入词嵌入 w 和需要查询的语义距离 d 后,云储存端返回相应的密文集合$\{C_{SP_{w,d}*}\}$。为了避免敏感信息的泄露,词嵌入 $SP_{w,d}*$ 集合的每个词嵌入由客户端生成搜索陷门提交云存储端,完成搜索请求。类似地,如果搜索是词嵌入集合 $W=\{w_1,w_2,\cdots,w_q\}$和语义距离 d,则返回的密文集合为$\{C_{SPW,d}*\}$。

在索引建立阶段,客户端随机选择大素数 a 和 b 作为私钥,并把每一个词嵌入 w 建立索引 $I_w=I(T_w,Enc(sk_w,FID_w))$,其中 $T_w=f(a,w)$,$sk_w=g(b,w)$;同时将索引 I 和密文 $C=Enc(x_{\text{key}},F)$发送到云存储端进行存储。

云存储端在接收到客户端的搜索陷门 $T_{SP_{w,d}*}$ 时,执行索引搜索操作 $S_{SP_{w,d}}=I'(T_{SP_{w,d}}*)$,并将满足条件的索引集合 CID'和对应密文C'返回给客户端。

客户端在收到 CID'和 C',授权用户使用密钥对密文 C'进行解密操作 $F'=Eec(x_{\text{key}},C')$,对 $FID'=Dec(sk_w,CID')$解密密文 ID,其中 $sk_w=g(b,w)$,最后验证 FID'和 F'的正确性。

5 安全性分析

文档及索引的隐私性:在选择对文档及索引的加密操作时,任何传统的加密方式都可行;而且文档与索引的加密方式可以是不同的加密算法,这不影响搜索的正确性,并保证了加密算法不易被攻击。

模式的安全性:当用户输入相同的词嵌入和语义距离时,云储存端总是会返回相同的结果来保证攻击者不能通过查询模式来分析出有价值的信息。服务器能得到的安全视图仅包括加密文件 C、加密索引文件 CID 和词嵌入陷门 $T_{SP_{w,d}*}$,此搜索请求是可以接受的,云存储端不能根据视图信息追踪到明文或语义信息。

陷门隐私性:由于词嵌入的陷门生成的是伪随机的单向函数,对于用户查询同一关键词或关键词集合时生成的是不同的陷门,服务器无法区分相同词嵌入的搜索凭证。

明文与明文索引校验：为了防止云存储端的加密文件 C 和加密索引文件 CID 被恶意篡改或删除，客户端在收到 C' 和 CID' 时，分别解密得到 FID' 和 F'，如果 FID' 和 F' 保持一致则表示正确，保证了数据的正确性和完整性。

6 实验分析

本研究提出的基于词嵌入的云存储可以搜索采用的具体方案为全同态加密的方案[15]，全同态加密有别于其他加密方法，包括 4 个操作：密钥生成算法、加密算法、解密算法和密文计算。其中密文计算有别于其他加密算法的重要内容；密文计算用于 $S_{SP_{w,d}}=I'(T_{SP_{w,d}}*)$ 的计算中，返回的结果保证了密文的正确性。文献[16]利用同态特性提出了多方排序方案，可用于云检索。在云安全存储方案中，李美云等人[17]利用加法和乘法同态实现了对密文的检索、存储问题。

6.1 时间复杂度

完成词嵌入的建立到搜索过程，一共要经历 3 个阶段：1）初始化词嵌入的过程。本研究采用 Word2vec 方案建立每个词嵌入，如果构建 D 维的词嵌入，时间复杂度为 $o(|D|\times\log_2(|W|))$。2）词嵌入索引建立的时间复杂度仅与词的个数 W 与文档数 M 有关，即建立所有的词嵌入索引需要遍历所有的文档所有时间复杂度为 $o(|W|\times|M|)$。3）在搜索过程中的时间复杂度上，用户根据关键词 w 与语义距离 d 得到的 $SP_{w,d}*$，完成一次搜索的时间复杂度为计算所有索引的时间，即 $o(|SP_{w,d}*|\times|W|)$。由于词嵌入与索引的建立都是离线处理的，所以仅考虑搜索的时间复杂度 $o(|SP_{w,d}*|\times|W|)$。

6.2 实验对比

本研究以维基百科中文文档[18]作为搜索和验证对象进行实验，首先利用 jieba 分词工具获得词条 716 091 个，分别对每个词建立 50，100，200 维词嵌入做比对实验，之后构建相应的索引。在完成这两项操作后，分别进行了 2 项实验，词嵌入的搜索时间消耗如图 2 所示。搜索精确率如图 3 所示。

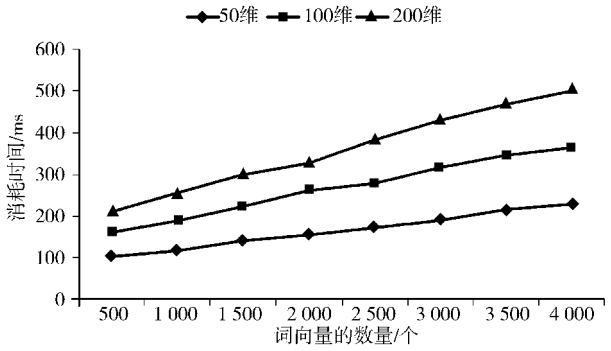

图 2　不同维度词嵌入查询时间消耗

Fig. 2　Query time of different dimension word embedding

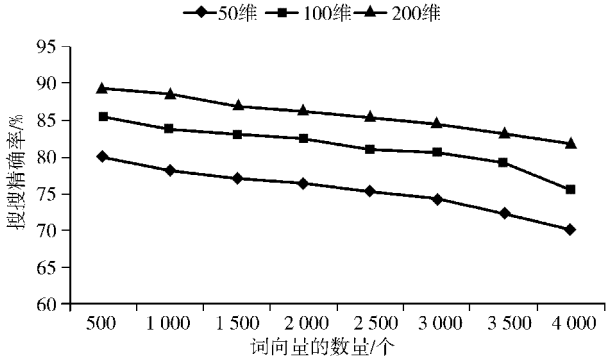

图 3　不同维度词嵌入的返回精确率

Fig. 3　Accuracy of different dimension word embedding

从以上实验结果可以看出，基于词嵌入的可搜索加密方案可行。搜索结果与词嵌入的集合与维度成线性增长，维度的增加提高了搜索的精确率；与此同时也增加了云端服务器的负担，消耗时间也随之增加。

7 结束语

本研究提出基于高维词嵌入的可搜索加密方案。该方案通过构建词嵌入及相关语义的词嵌入合集解决加密语义搜索问题。实验证明了方案的有效性和可行性，解决了用户终端的计算开销问题。采用全同态加密算法来实现整个过程，结果表明维度越高的词嵌入搜索精确率越高，同时开销也随之增大。在未来的工作中，将考虑返回索引相关度的排序问题，以提高返回精度。

参考文献：

[1] WEBER T.Cloud computing after Amazon and Sony: read for primetime[EB/OL].(2011-11-2)[2016-12-23].http://www.bbc.co.uk/news/business-13451990.

[2] 谢茂森，杨青.模糊综合评价在信息系统安全等级定级中的应用[J].重庆师范大学学报(自然科学版)，2014，31(5)：89-94.
XIE M S，YANG Q.Application of fuzzy comprehensive evaluation on the level of security classification information system[J].Journal of Chongqing Normal University (Natural Science)，2014，31(5)：89-94.

[3] 胡向东，刘竹林.网络化测控系统的信息安全方法研究[J].重庆理工大学学报(自然科学)，2016，30(5)：81-87.
HU X D，LIU Z L.Research on methods of information se-

curity for networked measurement and control systems[J]. Journal of Chongqing University of Technology (Natural Science),2016,30(5):81-87.

[4] 黄永峰,张永岭,李星.云存储应用中的加密存储及其检索技术[J].中兴通信技术,2010,16(4):33-35.
HUANG Y F,ZHANG Y L,LI X.Encrypted storage and its retrieval in cloud storage applications[J].ZTE Communications,2010,16(4):33-35.

[5] 沈志荣,薛巍,舒继武.可搜索加密机制研究与进展[J].软件学报,2014,25(4):880-895.
SHEN Z R,XUE W,SHU J W.Survey on the research and development of searchable enryption schemes[J].Journal of Software,2014,25(4):880-895.

[6] 陈燕俐,杨华山.可支持属性撤销的基于 CP-ABE 可搜索加密方案[J].重庆邮电大学学报(自然科学版),2016,28(4):545-554.
CHEN Y L,YANG H S.CP-ABE based searchable encryption with attribute revocation[J].Journal of Chongqing University of Posts and Telecommunications (Natural Science Edition),2016,28(4):545-554.

[7] SONG D,WAGNER D,PERRIG A.Practical techniques for searches on encrypted data[C]//Proc of the 2000 IEEE symp on security and privacy.Berkeley:IEEE Computer Society,2000.

[8] GOH E J.Secure indexes[EB/OL].(2003-04-01)[2016-12-23].http://eprint.iacr.org/2003/216.pdf.

[9] BONEH D,CRESCENZO G,OSTROVSKY R,et al.Public key encryption with keyword search[C]//Advances in cryptology-EUROCRYPT 2004.Tnterlaken:Springlink,2004.

[10] CURTMOLA R,GARAY J,KAMARA S,et al.Searchable symmetric encryption:improved definitions and efficient constructions[C]//Proc of the 13th ACM conf on computer and communications security(CCS).New York:ACM Press,2006.

[11] 吴继康,于徐红.混合云环境中多用户数据共享问题研究[J].计算机应用研究,2016,33(11):3435-3441.
WU J K,YU X H.Research on multi-user data sharing in hybrid cloud environment[J].Application Research of Computers,2016,33(11):3435-3441.

[12] MIKOLOV T,CHEN K,CORRADO G,et al.Distributed representations of words and phrases and their compositionality[J].Nips,2013:1-9.

[13] 王尚平,刘利军,张亚玲.可验证的基于词典的可搜索加密方案[J].软件学报,2016,27(5):1301-1308.
WANG S P,LIU L J,ZHANG Y L.Verifiable dictionary-based searchable encryption scheme[J].Journal of Software,2016,27(5):1301-1308.

[14] MIKOLOV T,CHEN K,CORRADO G,et al.Efficient estimation of word representations in vector space[J].Arxiv,2013,6:1-12.

[15] 刘明洁,王安.全同态加密研究动态及其应用概述[J].计算机研究与发展,2014,51(12):2593-2603.
LIU M J,WANG A.Fully homomorphic encryption and its applications[J].Journal of Computer Research and Development,2014,51(12):2593-2603.

[16] 肖倩,罗守删,陈萍.半诚实模型下安全多方排序问题的研究[J].电子学报,2008,36(4):709-714.
XIAO Q,LUO S S,CHEN P.Research on the problem of secure multi-party ranking under semi-honest model[J].Acta Electronica Sinica,2008,36(4):709-714.

[17] 李美云,李剑,黄超.基于同态加密的可信云存储平台[J].信息网络安全,2012(9):35-40.
LI M Y,LI J,HUANG C.A credible cloud storage platform based on homomorphic encryption[J].Netinfo Security,2012(9):35-40.

[18] WIKIPEDIA.Wikimedia Chinese dumps corpus[EB/OL].(2016-11-01)[2016-12-23].https://dumps.wikimedia.org/zhwiki/.

Searchable Encryptionscheme of Cloud Storage Based on Word Embedding

ZHONG Han[1], GUO Fei[2]

(1. Department of Computer Science & Technology, Nanjing University, Nanjing 210064;
2. College of Computer Science, Chongqing University, Chongqing 400033, China)

Abstract: [**Purposes**]Cloud storage data security attracts more attention, these are difficult problems of how to manage the encrypted data under the premise of ensuring the security of cloud storage data, and how to retrieval the encrypted data. A searchable encryption scheme is proposed for cloud storage based on word embedding. [**Methods**]This scheme establishes the high-dimensional word embedding for building security indexing, and increases the semantic distance to extend the keywords set. The pseudo-random function is constructed to secure the private key and keywords. [**Findings**]This design is to ensure the security of data storage, and also to provide the flexible of retrieval, which avoids the leakage of key information during user retrieval. [**Conclusion**]The scheme is verified by using fully homomorphic encryption, and the results on the Wikipedia dataset testing show that the higher the dimension word embedding, the higher the precision of search, meanwhile the overhead of cloud storage is increased.

Keywords: cloud storage; searchable encryption; word embedding; fully homomorphic

(责任编辑 游中胜)