基于 RFCA 的概念相似度计算方法*)

曹泽文　陆昌辉　张维明　邓　苏
(国防科学技术大学信息系统与管理学院　长沙 410073)

摘　要　本文主要针对基于 FCA 的概念相似度计算模型不能计算非形式概念之间的相似度问题，引入粗糙集理论，给出了 FCA 中等价关系的定义，提出了基于 RFCA 的概念相似度计算方法，通过实例说明该方法是可行的。

关键词　FCA，Rough Set，RFCA，概念相似度

A RFCA-based Approach for Concept Similarity Computation

CAO Ze-Wen　LU Chang-Hui　ZHANG Wei-Ming　DENG Su
(College of Information System & Management, NUDT, Changsha 410073)

Abstract　In order to solve the problem the FCA-based concept similarity computational model can't compute similarity of two informal concepts, rough set theory is then introduced, the definition of equivalence relation of concept lattice is given, a novel similarity measure method based on rough set and formal concept analysis (RFCA) is proposed. The example proves the method is feasible.

Keywords　FCA, Rough set, RFCA, Concept similarity

1　引言

随着本体应用的增多，如何解决异构本体间的互操作已成为一个棘手的问题[1]。本体映射能很好地解决本体异构问题，是发现两个相同领域本体的概念之间的相关性(映射关系)的过程，是本体间概念和关系取得一致性的一个规范说明。本体映射也是本体结盟、本体集成、本体合并、本体翻译等的技术基础[2~4]。

本体一般可理解为概念、属性和关系的集合。属性即概念的属性，关系即概念间的关系。因此，本体映射的核心内容是计算两个概念的相似度，得到概念的相似矩阵。当其相似度大于某个阈值时，就认为这两个概念之间存在一定的映射关系。

目前已提出了多种概念相似度计算模型。从广义上可以将它们分成两组：连续度量空间模型、集合理论匹配模型[5]。前者的典型例子有 Shepard 模型，它是基于概率分布的。后者可以进一步分成：几何(Geometric)模型、转换(Transformational)模型、特征(Featural)模型、基于结盟(alignment-based)的模型等。几何模型是在 n 维空间计算实体特征向量之间的距离来判断概念之间的相似程度。转换模型是通过将一个实体转换为另一个实体所需要的转换步数来判断概念之间的相似程度。例如，DNA 序列 ACCG 要转换成 ACGA，需要两次转换过程。特征模型是考虑实体的公共特征集合的数量来判断概念之间的相似程度，其一个典型例子是 Tversky 提出的比率模型[6]。Rodriguez 和 Egenhofer 在 Tversky 模型的基础上，提出了一个利用特征数量计算两个本体中概念相似度的模型[7]，但是 Rodriguez 和 Egenhofer 模型仅仅单纯地使用相同特征、不同特征的数量作为判断概念相似度的依据，没有考虑特征之间的结构关系，导致概念相似度计算结果准确度不高。

在 Rodriguez 和 Egenhofer 模型的基础上，Souza 和 Davis 将形式概念分析(Formal Concept Analysis，FCA)理论引入到本体映射过程中，提出基于 FCA 的概念相似度计算模型，以形式概念的不可约下确界元素作为相似度计算的依据，提高了概念匹配的准确率[8]。

但是，基于 FCA 的概念相似度计算模型只能计算形式概念对之间的相似度，对于形式背景中非形式概念(粗糙概念)之间的相似度计算就无能为力。为此我们将粗糙集理论引入到 FCA 中，提出基于粗糙形式概念分析(Rough Formal Concept Analysis，RFCA)的概念相似度计算模型，可以有效地解决粗糙概念之间的相似度计算问题。

本文首先简单介绍 FCA 与粗糙集理论的基本概念，然后分析基于 FCA 的概念相似度计算过程及存在的问题，给出了 FCA 中等价关系的定义，最后提出了基于 RFCA 的概念相似度计算模型，给出了模型计算实例。

2　形式概念分析 FCA 与粗糙集理论 RS

2.1　形式背景、形式概念与概念格

在哲学中，概念被理解为由外延和内涵两个部分所组成的思想单元。基于对概念的这种哲学理解，德国 R. Wille 教授于 1982 年首先提出了形式概念分析理论[9]，下面简单介绍相关概念。

定义 1　一个形式背景(Formal Context)是一个三元组 $K=(O,A,R)$，其中 O 是对象的集合，A 是属性的集合，R 是 O 和 A 之间的一个二元关系，即，$R\subseteq O\times A$。

一个小的形式背景能够用一个矩形表来表示，表的每一行是一个对象，每一列是一个属性。若 g 行 m 列的交叉处是 x，则表示对象 g 具有属性 m。

*)国家自然科学基金(60172012)、武器装备预研基金资助项目(51421020904KG01)。曹泽文　副教授，博士生，主研方向为知识系统、决策支持系统等。

定义 2 对于一个对象集合 $E\subseteq O$，我们定义：$E'=\{m\in A\mid \forall g\in O, gRm\}$；对于一个属性集合 $I\subseteq A$，我们定义：$I'=\{g\in O\mid \forall m\in I, gRm\}$。

定义 3 形式背景 (O,A,R) 的一个形式概念(Formal Concept)是 (E,I)，其中 $E\in O, I\in A$，且满足 $E'=I$ 和 $I'=E$。我们称 E 是形式概念 (E,I) 的外延，I 是形式概念 (E,I) 的内涵。$\underline{\mathfrak{R}}(O,A,R)$ 表示形式背景 (O,A,R) 的所有形式概念的集合。

形式概念 (E,I) 的定义意味着：E 是共享所有 I 中相同属性的所有对象构成的集合，I 是被对象集 E 中相同对象共享的所有属性的集合。E' 和 I' 的定义以及约束 $E'=I$ 和 $I'=E$，使得 E 和 I 之间建立 Galois 连接的充要条件得到了满足。

定义 4 如果定义在集合 H 上的一个二元关系 R，对于任意元素 $x,y,z\in H$，都满足下列条件的话，那么我们称 R 是一个偏序关系(或者简称为序)：

(1) xRx (自反)

(2) xRy 且 $yRx\Rightarrow x=y$ (反对称)

(3) xRy 且 $yRz\Rightarrow xRz$ (传递)

对于偏序关系 R，我们经常使用符号 $\leqslant$ 来表示(对于 R^{-1}，使用符号 $\geqslant$ 来表示)，如果 $x\leqslant y$ 而且 $x\neq y$，则我们写作 $x<y$。通常，我们读作 $x\leqslant y$ 为“x 小于或等于 y”。偏序集 $\langle H,\leqslant\rangle$ 是一个有序二元组，它由一个集合 H 与定义在 H 的一个偏序关系 $\leqslant$ 组成。

定义 5 如果 $X_1=(E_1,I_1)$ 和 $X_2=(E_2,I_2)$ 是一个形式背景的两个形式概念，而且 $E_1\subseteq E_2$、$I_2\subseteq I_1$，那么我们称 X_1 是 X_2 的子概念(subconcept)，X_2 是 X_1 的超概念(superconcept)，并记作 $X_1\leqslant X_2$，或 $(E_1,I_1)\leqslant(E_2,I_2)$。关系 $\leqslant$ 是定义在形式概念上的偏序(或者简称为序)。

进一步考察形式概念以及存在的偏序关系，还可以发现另外的属性：对每个非空的形式概念集总是存在一个唯一的最大子概念(meet，也叫下确界)和一个唯一的最小超概念(join，也叫上确界)。

偏序集 $(H,\leqslant)$ 加上它所具有的属性构成一种新的数据结构：概念格。

定义 6 对于形式背景 $H=(O,A,R)$，存在唯一的一个偏序集 $\langle H,\leqslant\rangle$ 与之对应，并且该偏序集存在一个唯一的下确界和一个唯一的上确界，这个偏序集产生的格结构称为概念格(concept lattice)，记为 $\underline{\mathfrak{R}}(O,A,R)$。

定理 1 概念格 $\underline{\mathfrak{R}}(O,A,R)$ 是一个完全格，其任意形式概念元素 (E_t,I_t) 的下确界 $\wedge$ 和上确界 $\vee$ 分别如下式所示：

$$\wedge(E_t,I_t)=(\bigcap E_t,(\bigcup I_t)'')$$

$$\vee(E_t,I_t)=((\bigcup E_t)'',\bigcap I_t)$$

定义 7 如果概念格中一个元素不能写成其它元素的下确界，则称该元素是不可约下确界(meet-irreducible)元素。如果概念格中一个元素不能写成其它元素的上确界，则称该元素是不可约上确界(join-irreducible)元素。

有一个直观的方法判断概念格中元素是否是不可约上、下确界元素，即在概念格中，如果元素向上只有一条边相连，则该元素是不可约上确界；如果向下只有一条边相连，则该元素是不可约下确界。

2.2 粗糙集理论简介

设 R 是论域 U 上的等价关系(满足自反性、对称性和传递性)，记为 $R\subseteq U\times U$。U 中所有与 x 具有等价关系 R 的元素的集合记为 $[x]_R=\{y\in U\mid (x,y)\in R\}$。$[x]_R(x\in U)$ 为最小的非空可定义集。将 $\bigcup\limits_{x\in U}[x]_R$ 也作为可定义集，等价划分 U/R 构成的 σ 一代数称为可定义集的全体，记为 $\sigma(U/R)$。若 $X\subseteq U$，但 $X\notin\sigma(U/R)$，则称为不可定义的。

商集 $U/R=\{[x]_R\mid x\in U\}$ 是等价关系 R 将论域 U 进行划分所得的等价类的集合。给定 $X\subset U$，要用 U/R 中的元素来描述、表达 X，不一定能精确地进行。但常常可以用关于 X 的一对下近似、上近似来界定 X，这导致粗糙集概念的产生。

定义 8 设 R 是论域 U 上的等价关系，对象集 $X\subseteq U$，$(\underline{R}X,\overline{R}X)$ 称为在 Pawlak 近似空间 (U,R) 上的一个粗糙近似，其中

$$\underline{R}X=\{x\in U\mid [x]_R\subseteq X\}$$

$$\overline{R}X=\{x\in U\mid [x]_R\cap X\neq\phi\}$$

$\underline{R}X$、$\overline{R}X$ 分别称为 R 的下近似和 R 的上近似。若 $\underline{R}X\neq\overline{R}X$，则称 X 为 R 粗糙集；否则 X 为 R 可定义集。

以上是经典的 Pawlak 意义下的粗糙集概念[10]。

3 基于 FCA 的概念相似度计算模型

目前，FCA 技术已成功应用于数据挖掘、软件工程、信息检索、知识库组织、本体合并等诸多领域[11,12]。本文主要考虑本体结盟过程中通过 FCA 技术建立概念之间的语义相似度。具体实现方法是：建立本体结盟后的形式背景，以此为基础形成概念格；引入新的基于概念格的相似度计算模型，以形式概念的不可约下确界元素作为相似度计算的依据，计算形式概念之间的语义相似度。

3.1 建立形式背景与相应概念格

下面以文[12]中给出的两个本体：生产厂家(Factory)与运输公司(Carrier)所建立的交通工具 Transportation 本体片断示例，来说明基于 FCA 的概念相似度计算方法。

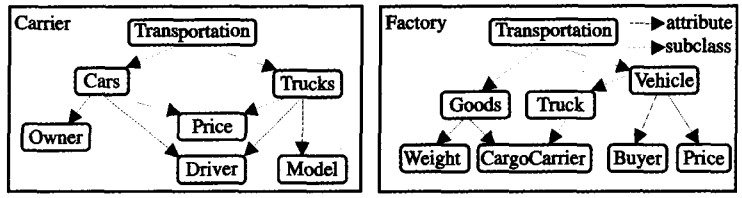

图 1 生产厂家与运输公司中交通工具本体片断的示例

依据文[8]得到本体 Carrier 和 Factory 连接后的上层本体的形式背景，如表 1 所示。表的行对应对象，列对应属性。如果一个对象具有某一属性，则在对象与属性的交叉位置处标上“×”。Factory 本体中的对象名称前标注 A，Carrier 本体中的对象名称前标注 B。

利用文[14]中的软件 ConExp 形成形式背景的相应概念格，即得到相应 Hasse 图，如图 2 所示。结点上的名字表示属性，结点文本框中的名字表示本体 Carrier 和 Factory 中对

象。标注在概念格某一结点上的对象继承该结点到根结点路径上的所有结点的属性。例如，对象 A CargoCarrier 所在结点到根结点路径为：11-9-7-3-0，因此对象 A CargoCarrier 具有属性 truck、motorVehicle、vehicle、transport 和 artifact。

表 1　本体 Carrier 和 Factory 连接后的上层本体的形式背景

	transport	driver	person	vehicle	car	truck	motorVehicle	buyer	owner	price	goods	artifact
A transportation	X											X
A vehicle	X		X	X				X		X		X
A buyer			X					X				
A price										X		
A truck	X			X	X		X					X
A cargo carrier	X			X	X		X					X
A goods	X			X	X		X				X	X
B transpor	X											X
B cars	X	X	X	X		X	X		X			X
B trucks	X	X	X	X	X		X					X
B driver		X	X									
B owner			X						X			
B price										X		

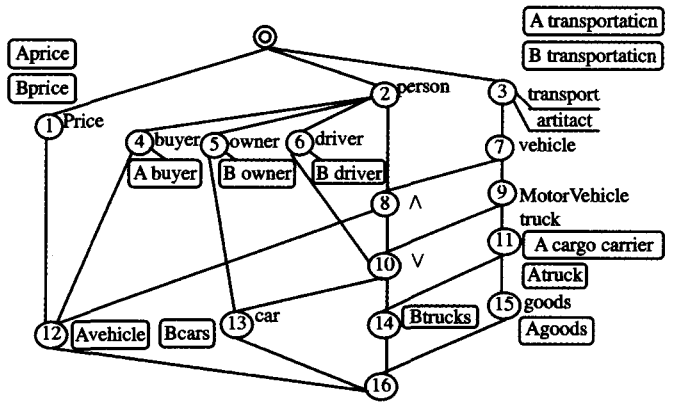

图 2　与表 1 对应的形式背景的 Hasse 图

3.2　基于概念格计算概念之间的相似度

概念格中的元素(结点)可以分成可约下确界元素与不可约下确界元素，其中可约下确界元素可以通过其它元素生成，也就是说，它的引入没有向概念格中增加新的信息。相反，不可约下确界元素是概念格结构所必需的。为此，Souza 和 Davis 在 Rodriguez 和 Egenhofer 特征相似度模型的基础上，将对象集的公共特征集换成公共的不可约下确界元素集，得到了一个基于 FCA 的概念相似度计算模型：

$$Sim(a,b)=\frac{|(a\vee b)^{\wedge}|}{|(a\vee b)^{\wedge}|+\alpha(a,b)|(a-b)^{\wedge}|+(1-\alpha(a,b))|(b-a)^{\wedge}|} \tag{2}$$

其中$(a\vee b)$表示计算形式概念 a 和 b 的上确界，$(a\vee b)^{\wedge}$表示形式概念 a 和 b 的上确界的内涵中的不可约下确界元素集。$(a-b)^{\wedge}$表示在 a 中、不在 b 中的不可约下确界元素集；$(b-a)^{\wedge}$表示在 b 中、不在 a 中的不可约下确界元素集。$\alpha(a,b)=0.5$，也就是说，基于概念格的相似度计算模型中，相似度是对称的，即如果 a 与 b 相似，则 b 与 a 具有同样的相似度。

例如，假设要计算形式概念 $a=12$ 和 $b=14$ 的相似度，a 与 b 相应的对象分别是 A vehicle 和 B trucks。

$(a\vee b)^{\wedge}=\{2,3,7\}$

$(a-b)^{\wedge}=\{1,2,3,4,7\}-\{2,3,6,7,9,11\}=\{1,4\}$

$(b-a)^{\wedge}=\{2,3,6,7,9,11\}-\{1,2,3,4,7\}=\{6,9,11\}$

$$Sim(a,b)=\frac{3}{3+0.5*2+0.5*3}=0.545$$

因此，概念 A vehicle 与 B trucks 的语义相似度为 0.545。

3.3　问题分析

基于 FCA 的本体映射方法中只能计算形式概念对之间的相似度，即要求形式概念(E,A)满足：$E\subseteq G, A\subseteq M$，且 $E'=A, A'=E$。对于形式背景 $K(G,M,I)$ 中任意概念(E,A)，其中 $E\subseteq G, A\subseteq M$，但不满足条件：$E'=A, A'=E$，因此该概念不是一个形式概念，在 Hassse 图上没有对应的结点，也没有不可约下确界元素。计算这类概念之间的相似度就不能直接采用公式(2)，为此我们引入粗糙集理论。

4　基于 RFCA 的概念相似度计算方法

4.1　FCA 中等价关系

在形式背景 $K(G,M,I)$ 下，可以根据关系 I 确定对象和属性的等价类。

对象等价类：设 $gI=\{m\in M|gIm\}$ 为对象 g 具有的所有属性的集合，定义一个 R 在 G 上的关系：g_1Rg_2，当且仅当 $g_1I=g_2I$。这里 $g_1, g_2\in G$。G/R 表示 R 在 G 导出的所有等价类的集合。G/R 的等价类被称作基本集(elementary sets)。任何有限的基本集的并被称作是一个可定义(definable)集。

属性等价类：令 $Im=\{g\in G|gIm\}$ 为属性 m 拥有的所有对象的集合，定义一个 R' 在 G 上的关系如下：$m_1R'm_2$ 当且仅当 $Im_1=Im_2$，这里 $m_1, m_2\in M$，M/R' 表示 R' 在 M 导出的所有等价类的集合。

例如，对上节 carrier-factory 结盟本体对应形式背景 $K(G,M,I)$，$G/I=\{\{A\ price, B\ price\}, \{A\ transportation, B\ transportation\}, \{A\ buyer\}, \{B\ owner\}, \{B\ driver\}, \{A\ vehicle\}, \{B\ cars\}, \{B\ trucks\}, \{A\ goods\}, \{A\ truck, A\ cargo\ carrier\}\}$，$M/I=\{\{price\}, \{person\}, \{transport, artifact\}, \{buyer\}, \{owner\}, \{driver\}, \{vehicle\}, \{motorvehicle\}, \{truck\}, \{goods\}, \{car\}\}$。

4.2 粗糙概念之间的相似度计算模型

由于已在 FCA 中定义了等价类，这样就可以研究概念的可定义问题。概念可以分为两类：可定义概念、不可定义概念。

定义 9 在一个形式背景 $K(G,M,I)$ 上的对 (E,A) 被称为概念，$E\subseteq G, A\subseteq M$。若 (E,A) 是一个形式概念，则称 (E,A) 是可定义概念，否则为不可定义概念。两个概念相等是指有相同的属性和对象。

定义 10 对于形式背景 $K(G,M,I)$ 上的对象集和属性集 $E\subseteq G, A\subseteq I$，为关系 R 上的粗糙集，称 (E,A) 为粗糙概念，$((\underline{R}E')'\cap(\overline{R}A)', ((\underline{R}E')'\cap(\overline{R}A)')')$ 为概念 (E,A) 的下近似，记为 $(E,A)_-=((\underline{R}E')'\cap(\overline{R}A)', ((\underline{R}E')'\cap(\overline{R}A)')')$。$(E,A)$ 的上近似 $(E,A)^-=((\overline{R}E)'\cap((\underline{R}A)')', ((\overline{R}E)'\cap(\underline{R}A)')')$。

而且，粗糙概念 (E,A) 的下近似、上近似都是形式概念[13]。因此，通过下近似、上近似运算，达到了将粗糙概念 (E,A) 转变成形式概念的目的，然后就可以利用基于 FCA 的概念相似度计算公式对粗糙概念进行相似度计算。因此，我们给出基于 RFCA 的概念相似度计算模型为：

$$Sim(a,b)=\frac{|(a_-\vee b_-)^\wedge|}{|(a_-\vee b_-)^\wedge|+\alpha(a,b)|(a_--b_-)^\wedge|+(1-\alpha(a,b))|(b_--a_-)^\wedge|} \quad (3)$$

其中 a、b 表示两个粗糙概念，a_-、b_- 分别表示概念 a 和 b 的下近似，$(a_-\vee b_-)$ 表示计算 a_- 和 b_- 的上确界，$(a_-\vee b_-)^\wedge$ 表示 a_- 和 b_- 的上确界的内涵中的不可约下确界元素集。$(a_--b_-)^\wedge$ 表示在 a_- 中、不在 b_- 中的不可约下确界元素集；$(b_--a_-)^\wedge$ 表示在 b_- 中、不在 a_- 中的不可约下确界元素集；$\alpha(a,b)=0.5$。

例如，对上节 carrier-factory 结盟本体对应形式背景 $K(G,M,I)$，给定一个粗糙概念 $a=(E,A)=(\{A\ truck, B\ trucks\}, \{vehicle, truck, artifact\})$。

$E=\{A\ truck, B\ trucks\}$，$\underline{R}E=\{B\ trucks\}$

$\underline{R}E'=\{transport, driver, person, vehicle, truck, motorvehicle, artifact\}$

$(\underline{R}E')'=\{B\ trucks\}$

$A=\{vehicle, truck, artifact\}$，$\underline{R}'A=\{vehicle, truck, artifact\}$

$\overline{R}'A=\{vehicle, truck, artifact, transport\}$

$(\underline{R}'A)'=\{A\ truck, B\ trucks, A\ cargo\ carrier, A\ goods\}$

$(\overline{R}'A)'=\{A\ truck, B\ trucks, A\ cargo\ carrier, A\ goods\}$

$(\underline{R}E')'\cap(\overline{R}'A)'=\{A\ truck, B\ trucks, A\ cargo\ carrier, A\ goods\}$

$(\underline{R}E')'\cap(\overline{R}'A)'=\{B\ trucks\}$

$((\underline{R}E')\cap(\underline{R}'A)')'=\{transport, driver, person, vehicle, truck, motorvehicle, artifact\}$

$a_-=(E,A)_-=(\{B\ trucks\}, \{transport, driver, person, vehicle, truck, motorvehicle, artifact\})$

此时 a_-，即 $(E,A)_-$ 已经是形式概念，具体对应概念格上的结点 14，可以对它进行不可约下确界元素集的计算。如果另外给定概念 $b=(\{A\ vehicle\}, \{transport, person, vehicle, buyer, price, artifact\})$，利用公式(3)计算 $Sim(a,b)=0.545$。

结束语 本文在基于 FCA 的概念相似度计算模型的基础上，通过引入粗糙集理论，提出了基于 RFCA 的概念相似度计算模型，有效地解决了粗糙概念之间相似度计算问题。形式概念分析与粗糙集理论作为两种数据分析和知识处理的形式化工具，已经被广泛地应用于软件工程、知识工程等领域，对两者的结合与应用是一个非常有前途的研究领域。

参 考 文 献

1 Doan A H. Learning to Map between Structured Representations of Data:[Ph thesis]. University of Washington,2002

2 Noy N, Musen M. PROMPT: Algorithm and Tool for Automated Ontology Merging and Alignment. In: Proc. AAAI2000. AAAI Press, 2000

3 Noy N F, Musen M A. SMART: Automated support for ontology merging and alignment [J]. In: Twelfth Workshop on Knowledge Acquisition, Moeling, and Management, Banff, Canada, 1999

4 Dou D, McDermott D, Qi P. Ontology Translation by ontology Merging and Automated Reasoning: [Ph thesis]. University of Yale, 2004

5 Tenenbaum J B, Griffiths T L. Generalization, similarity, and bayesian inference. Behavioral and Brain Sciences, 2001. 24

6 Rodriguez M A, Egenhofer M J. Determining semantic similarity among entity classes from different ontologies. IEEE Transactions on Knowledge and DataEngineering, 2003. 15

7 Tversky A. Features of Similarity. Psychological Review, 1977. 84

8 de Souza X S, Davis J. Aligning Ontologies and Evaluating Concept Similarities. Lecture Notes in Computer Science, Springer-Verlag, 3291, 2004

9 Ganter B, Wille R. Formal concept analysis: mathematical foundations [M]. Springer, Berlin: Springer-Heidelberg-New York, 1999

10 Pawlak Z. Rough set. International of Computer and Information Science, 1982, 11: 341～356

11 Stumme G, Maedche A. FCA-merge: Bottom-up merging of ontologies. In: Proc. 17th Intl. Conf. on Artificial Intelligence (IJCAI'01), Seattle, WA, USA, 2001

12 Mitra P, Wiederhold G, Kersten M L. Graph-Oriented Model for Articulation of Ontology Interdependencies. In: Proceedings of the International Conference on Extending Database Technology (EDBT). Volume 1777 of Lecture Notes in Computer Science (LNCS), Springer-Verlag, Konstanz, Germany, 2000

13 吴强，刘宗田. 在 FCA 中的粗糙概念. 小型微型计算机系统，2005，26(9)

14 Concept Explorer. http://www.sourceforge.net/projects/conexp