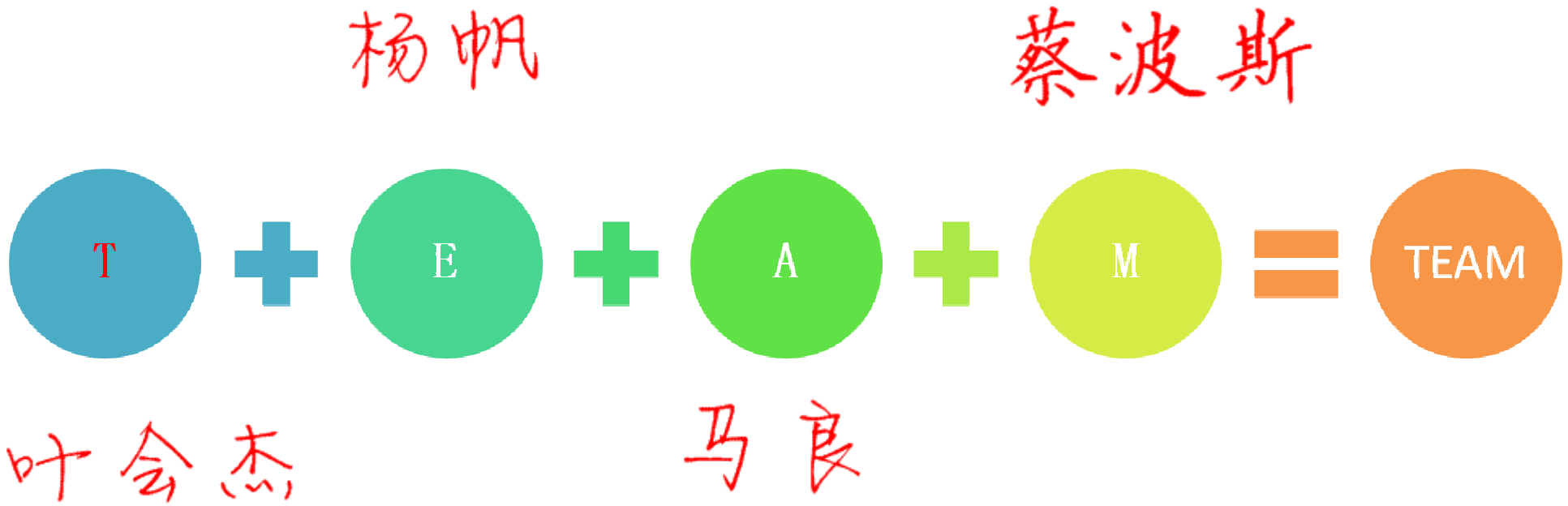
3.8语言解析

主讲：马良

指导老师：赵燕平 张华平

2010年11月13日


杨帆
蔡波斯
T
E
A
M
TEAM
叶会杰
马良

大纲

单个词

- 应用单个词的基础是扫描，为此需要解决几个技术上的问题——大小写的问题、词干还原的问题、停用词问题。

简单短语

- 单纯的使用一个词会产生那些问题呢?

复杂短语

- 它应用了NLP中的许多方法，但是这些方法似乎没有对整个的搜索有太大的贡献，因为它的研究才刚刚开始

单个词

- 搜索过程————检索词袋
- 问题：为文档选择合适的有区分度的关键词通常很困难
- 如果系统有分类，就更难将文档归入唯一的领域中了
- 搜索文档时需要人工干预，并为每一篇文档赋予一系列的相关词
- 人工对文档进行分类代价高昂

对文本的内容进行规范化

- 前缀、后缀、大小写匹配引起的问题
- Porter and loving 算法
- Red ring → r
- KSTEM方法，该方法利用词典以确保生成的每个词干都是有效的
- 停用词大概占文档集中的40%
- 去掉停用词可能会导致某些查询能力的下降
- 处理特殊字符的时候还需要用到其他的语言解析规则

方法很简单，而且很高效，他的效率经常不亚于许多复杂的算法

简单短语

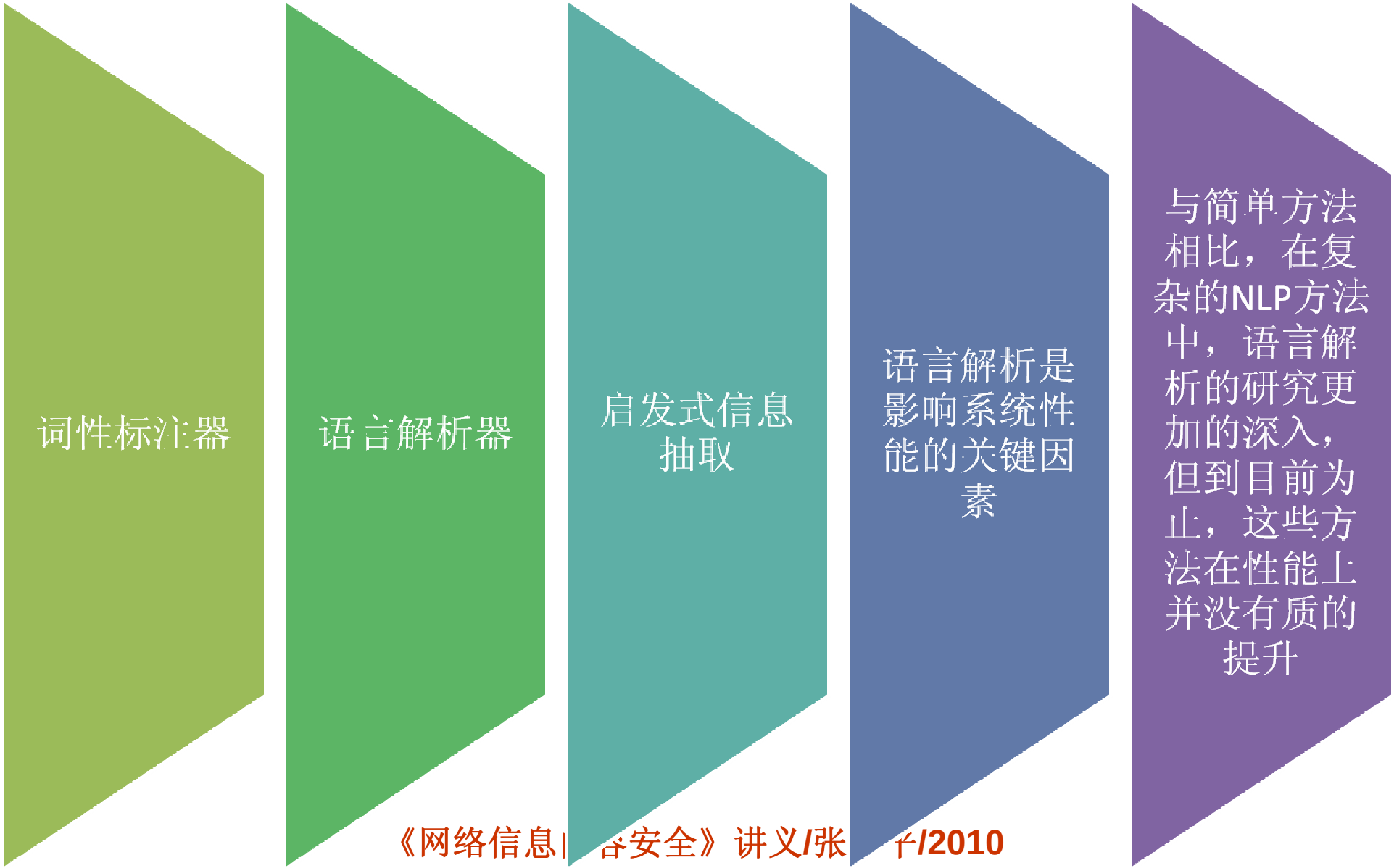
在文档检索中，使用短语来代替词元。

短语包含了词袋中所不具备的一些语义

NEW YORK

NOT NEW DUKE OF YORK

自然语言处理 NLP


词性标注器
语言解析器
启发式信息抽取
语言解析是影响系统性能的关键因素
与简单方法相比，在复杂的NLP方法中，语言解析的研究更加的深入，但到目前为止，这些方法在性能上并没有质的提升

NLP

20世纪60年代早期

NLP系统常常与信息检索系统是完全不同的

NLP系统试图通过建立一个表示文档的正则结构来理解文档的内容。

目的：减少语言本身固有的歧义

Walking： are moving slowly by gradually placing one foot in front of the other

NLP

- NLP系统用确定性的正则结构识别文档中的关键元素（主谓宾）
- Step1　语法分析
- Step2 建立单独的结构来表示文档
- 简单元语——包含一大类动词
- John drove to work
- John used his car to get to work
- John PTRANS work（physical transport）

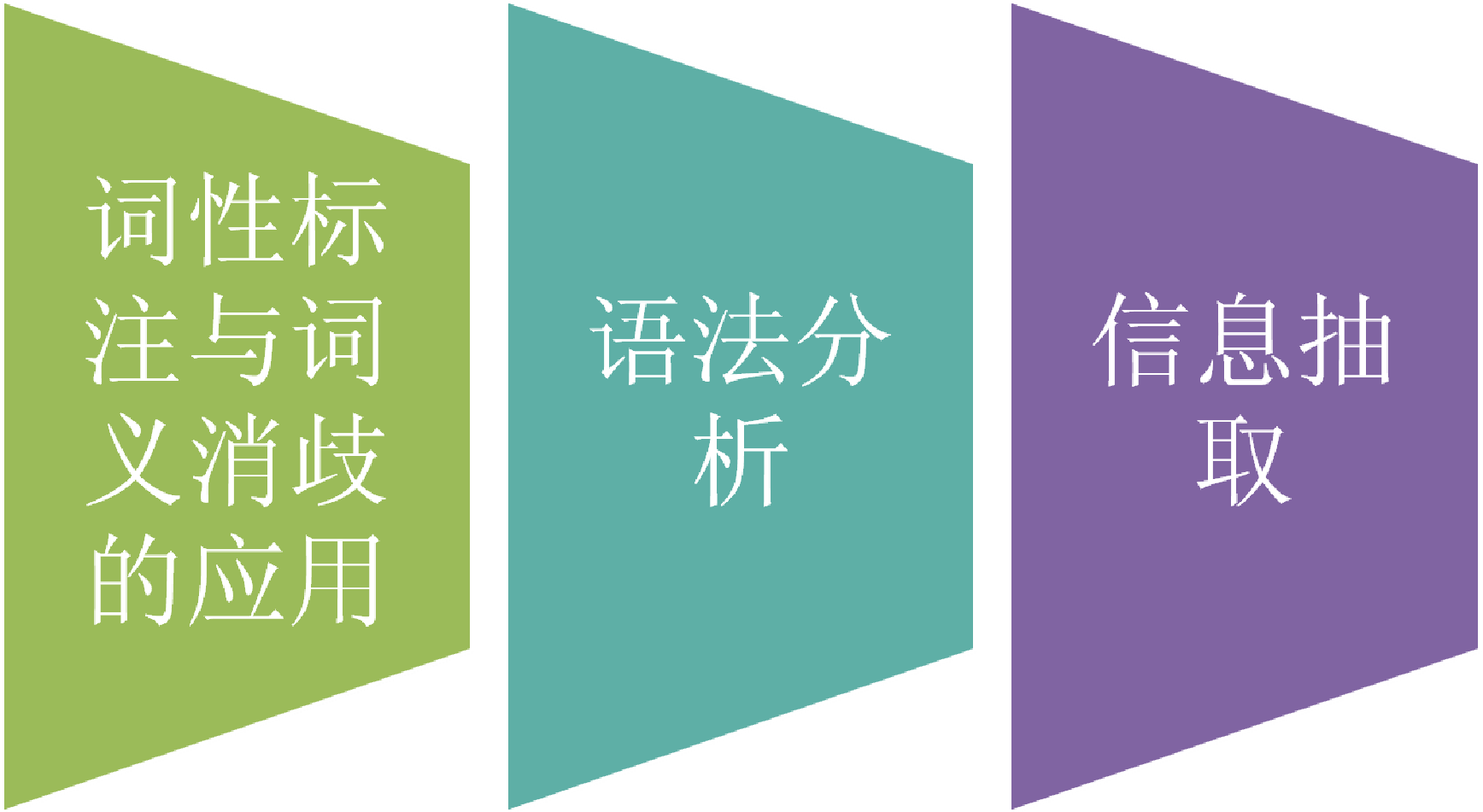
令人感到遗憾的NLP

- 尽管NLP处理技术取得了进步，但是知识表示中的许多问题导致构建一个满足需要的正则结构是极其困难的
- CYC在过去的15年人工创建了一个知识库，这个过程中很难确定大量文本中知识的准确含义

令人意想不到的NLP

- 许多在语言理解中效果不理想的工具在信息检索系统中却非常有用
- 词性标注器　语法分析器
- 识别文档关键词的几个算法

复杂短语



词性标注与词义消歧的应用

- 目的：切分文本，并确定切分出的符号串的词性
- 做法：使用已标记的语料库来统计下面两个量：
  - 一个给定词被赋予某个词性标记的频率
  - 不同词性标记序列出现的频率
- 有了这两类数据，可以使用动态规划的算法来优化某一处理步骤中符号串的词性标注

词性标注与词义消歧的应用

- 可以识别出名次序列已经名次前面的形容词序列
- 另一个作用是修改匹配过程
- 文档与查询中的词要匹配，并且词性也要匹配
- 尽管这在理论上是合理的但是却没有产生什么较好的效果

语法分析

- 目的：识别句子中的语法成分
- 一般情况下，语法分析在词性标注之后进行

- 两种方法：一个方法应用语法分析器
- 首先尝试使用增强转换网络，本质上就是一个非确定的有限状态状态自动机
- 主谓宾最为该网络的一个状态转换序列。

- 问题是复杂的句子通过自动机转换后会产生许多不同的路径，并且递归的调用有限状态自动机，因为他们包含的子结构也是主谓宾齐全的句子

语法分析

- 其他的分析算法Gomez的**WUP**（Word Usage Parser），使用词典查找每个词，然后每个词生成特定的状态序列。尽管他比ATN快，但是他需要大量的人工干预来构建用法词典
- 一些词法分析基于轻量级分析，该分析先进行一遍快速扫描与确定关键元素，然后使用规则方法，但是对于复杂句子，该分析器不能进行完整分析

语法分析的应用

- 语法分析后，检索系统就可以使用其生成的语法结构
- 一个简单的用法是，选择几个语法成分作为查询的有效成分来与文档进行匹配
- 用这种方法生成的短语能匹配英语中的许多变化
- American President 与President of America
- President who is in charge of America

信息抽取

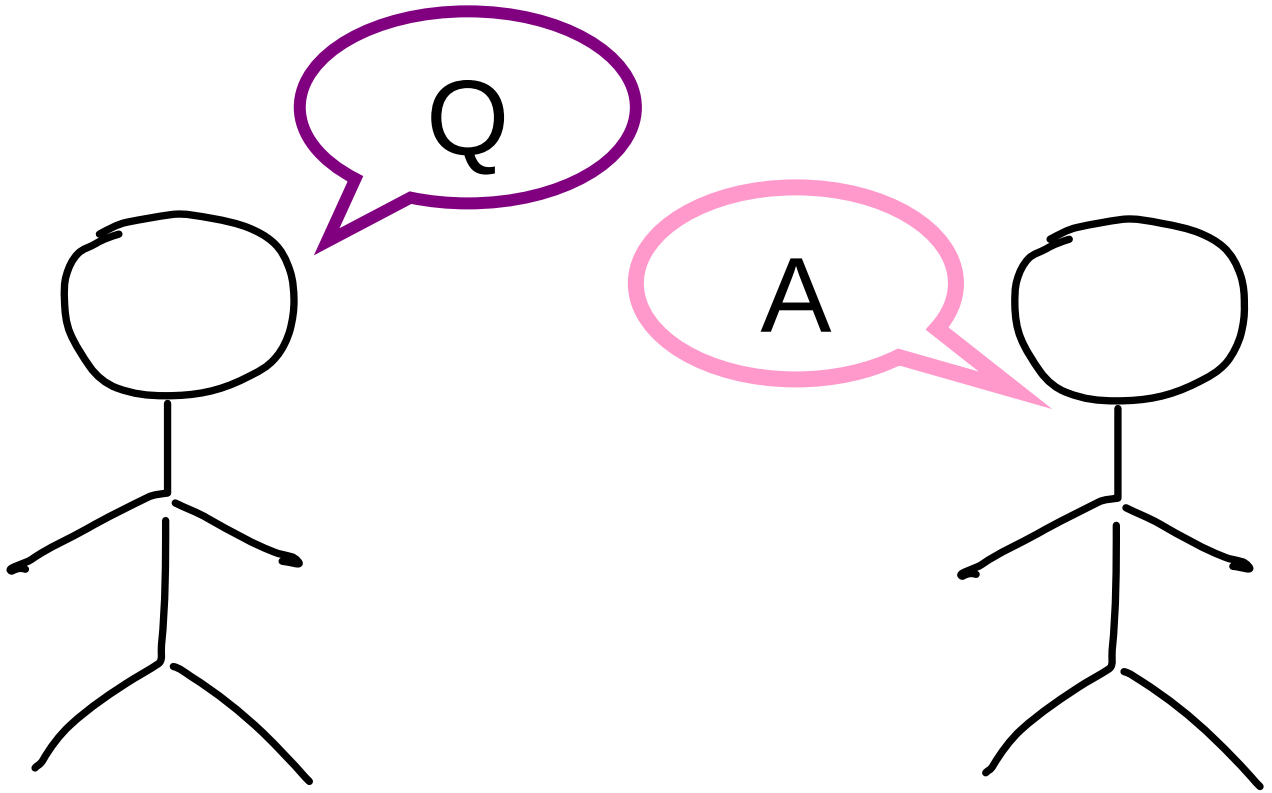
- 信息抽取——在非结构化的文档中查找各种结构化数据的方法。
- 主要是人名、地名、数量等的识别。
- 这些算法不是基于规则的，也不是基于统计算法，第一步大都是生成句子的语法结构，最少也是进行词性标注

信息抽取

- 命名实体识别器能识别人名、机构名以及地名
- 一个例子：用BBN公司的一个基于规则的的抽取程序来获得这篇新文档
- 规则人工制定
- 使用周围的词来确定该次是否应该被抽取

- <TEXT>
- Collins began his rise to success as the lightning-fingered guitarist for the jacksonville band formed in 1966 by a group of high school students.The band enjoyed national fame in the 1970’s with such hits as”Free Bird” ”Gimme Three Steps,” “Saturday Night Special” and Ronnie Van Zant’s feisty”Sweet Home Alabama”
- </TEXT>

- <TEXT>
- <ENAMEX TYPE="PERSON">Collins</ENAMEX> began his rise to success as the lightning-fingered guitarist for the <ENAMEX TYPE="lOCATION"> jacksonville </ENAMEX> band formed in <TIMEX TYPE="DATE">1966</TIMEX> by a group of high school students.The band enjoyed national fame in the <TIMEX TYPE="DATE"> 1970's </TIMEX> with such hits as"Free <ENAMEX TYPE="PERSON"> Bird </ENAMEX> ""Gimme Three Steps," "Saturday Night Special" and <ENAMEX TYPE="PERSON"> Ronnie Van Zant </ENAMEX> 's feisty"Sweet Home <ENAMEX TYPE="lOCATION"> Alabama </ENAMEX> "
- </TEXT>


Q
A

Thank you

Contact
Email: kevinzhang@bit.edu.cn

Welcome to visit my blog

http://hi.baidu.com/drkevinzhang/