从实际应用看 Google™ 语言模型的缺陷*

张化瑞

北京大学 计算语言学研究所，北京 100871

E-mail: hrzhang@pku.edu.cn

摘 要： 在 Google 应用中使用的语言模型与其公开的 Web 1T 5-gram 库有很大不同，一个根本的差异体现在是否忽略标点符号。本文以谷歌拼音输入法作为切入点，通过典型性实例揭示了 Google 语言模型的两个具有普遍性的缺陷：内嵌标点也算连续字串，外部链接视同文档内容。同时给出了弥补这些缺陷的建议。

关键词： Google 语言模型，缺陷，连续字串，标点符号，外部链接

Deficiency of Google™ Language Model Revealed by Real Applications*

Zhang Huarui

Institute of Computational Linguistics, Peking University, Beijing 100871

E-mail: hrzhang@pku.edu.cn

Abstract: The language model adapted in Google applications is quite different from the Web 1T 5-gram published by Google Inc. A fundamental disagreement lies in whether punctuations are omitted. Starting from Google Chinese Pinyin input method, two universal deficiencies of Google language model are revealed by typical examples: continuous string separated by punctuations, backlinks treated equally as page content. Advice to overcome these deficiencies is also provided.

Key words: Google language model, deficiency, continuous string, punctuation, backlink

1 目的和意义

Google 网页信息搜索在世界范围内处于领先地位，Google 统计机器翻译也多次在国际评测中独占鳌头，Google 的影响是有目共睹的，正因为如此，如果 Google 提供的服务在内在机制上有缺陷的话，那么对用户的影响也是不可忽视的。本文通过在实际应用中遇到的问题来分析 Google 语言模型可能存在的缺陷。

1.1 乘法效应

有人说，现在搜索引擎的用户满意度已经比较高了，根据 ACSI 的统计，大致在百分之七八十的样子，差不多了，还有必要在一些不影响大局的问题上细究吗？但事实上，影响搜索结果准确性的有很多因素。我们可以算一笔账：假如有十个独立的因素影响搜索的结

* 本文部分相关研究受到国家 973 项目资助（文本内容理解的数据基础，课题编号：2004CB318102）。

* Partially Supported by 973 Project, No. 2004CB318102.

果，对每个因素现在都能做到 70%的准确率，那么最后的准确率有多少？通过简单的计算不难得出：只有 3%左右(0.7^{10})。由此可见，不能满足于百分之七八十，如果在某个方面有一种统一而有效的方法能做到 100%，就不应该停留在 90%上，这样才能达到最终结果的不断完善。

1.2 长尾效应

从频度上来说，搜索结果十有七八符合我们的要求，但如果从效用上来看，又有多少呢？出现得多的容易搜得到、搜得准，但出现的少的是不是就没有用呢？事实上，恰恰相反，出现的少的很可能更有用一些，至少 Shannon(1948)的信息熵就从形式上说明了这一点，出现概率小的事件包括的信息量更大一些。Zipf(1935)早就发现语言中的词汇分布的统计规律：出现次数极多的数目很少，出现次数很少的数目很多，这就提示我们，在语言模型中，不能忽视在统计上不占优势的部分。Anderson(2006)提出，电子商务更应该重视为数众多的客户各种各样的个性化需求，而不仅仅是少数客户的批量订购，从而创造新的商业模式。

在学习和研究过程中，创新总是从尚未成为主流的地方开始，这样，如果那些刚刚萌芽、为数甚寡的新思路、新方法不能通过搜索有效获得的话，就有可能限制我们的视野，甚至使我们的工作成为重复劳动，这对创新是非常不利的。后面将通过一个实际的例子（“similarity entropy”）来说明。

1.3 对理性的尊重

有些问题比较复杂，不容易比出高低，比如搜索结果的排序；有些问题相对明晰，通过一定的逻辑分析，不难看出是否合理。比如，如果提问者想了解的是“虎”，而回答者提供的几乎全是关于“猫”的信息，提问者提出质疑，答复是：虎也是猫科，民间也有“猫”是“虎”的师傅的说法，因此，“猫”和“虎”的关系还是非常密切的，按照统一的综合多种因素的相关性排序，“猫”就排在了“虎”的前面，没有进行任何人工调整，所提供的结果是客观的。至于背后的原因，也许有猫比虎多得多（虎已是珍稀动物，猫却正大行其道）、与人的关系友好得多（猫是宠物，虎能伤人）等诸多因素，但提问者无法确定。不过提问者怎么也想不通：为什么问“虎”而答“猫”？

2 方法和数据

本文中采用的方法是基于实例的，而不是基于统计的。原因在于：对逻辑型的问题（值域为{0,1}），用典型的实例就能说明；对概率型的问题（值域为[0,1]），才需要引入大量的数据。

本文中的实例都是在实际应用中遇到的，或发现问题后针对问题进行扩展得到的，没有一个例子是虚构的。

本文中的例子均在 2008 年上半年内经过多次验证（比如，表 1 和图 2 即反映出两次搜索结果的不同），搜索结果有所变化，但其中包含的问题是一直存在的。

本文中涉及的谷歌拼音输入法的版本是 1.0.22.0. 需要说明的是，谷歌拼音输入法中的问题仅仅是发现 Google 语言模型的问题的一个切入点，该输入问题现已有所改进。

本文中的 Google 语言模型是指其功能的等效模型，并不涉及其内部实现，用黑箱方法研究其表现出来的功能。出于善意，我们首先相信 Google 在其网站上所作出的关于搜索技术和问题的说明，并由此推测其语言模型的问题所在，对原因的分析的可靠性依赖于 Google 说明的可靠性。但本文中所揭示的问题是确实存在的，而且是带有普遍性的，没有任何折扣。

本文中使用了谷歌拼音输入法和 Google 搜索的结果截图，Google 及相关图形标志是 Google Inc 的商标，在此特别声明。

在本文中，没有涉及 Google 的以下问题：

- 汉字编码的识别率问题；
- 中文分词的准确率问题。

因为这两个问题都不是逻辑型的，而是概率型的，随着技术的进步，性能会逐步提高。

3 内嵌标点也算连续字串？——“好山”“好水”不对称

在 Google 搜索中，如果要搜索一个完整的词组或连续字串，应该将其置于双引号内，曾经只支持英文的双引号，现在中文的双引号也可以。

下面使用以下符号约定：用方括号来标记中文输入的输入字符串或网页搜索的查询表达式，如 [hao] 表示输入“好”的全拼（3 个字母），[“号陕”] 表示在搜索框中输入双引号引起来的“号陕”两个字（共 4 个字符）。

3.1 问题的引出

在使用谷歌拼音输入法[6](1.0.22.0 版，该问题于 2007 年下半年发现)输入“好山好水好地方”这句话时，输入拼音[haoshan]后，排在首位的是“号陕”，而单独输入[hao]和[shan]排在首位的则是“好”和“山”，让人感到非常奇怪。因为“号陕”不是一个词，而且这两个字的组合以前就没怎么见过，所以觉得里面必有文章。

为什么“号陕”会排在首位？肯定是在 Google 的语言模型中认为这两个字经常在一起出现，所以才会出现没有“好山”只有“号陕”的情况。需要注意的是，单独输入[shan]时，“陕”连前五位都排不到，这样“号陕”这一组合就更值得注意。把[haoshanhaoshui]打完后，排在首位的是“号陕好水”，“好山”“好水”不对称。如图 1 所示。

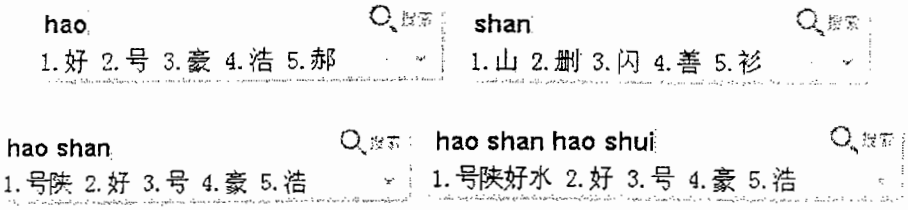


图 1 “好山好水”出“号陕”

3.2 “号陕”从何而来？

既然 Google 认为“号陕”是一个常用的组合，那就到 Google 搜索中搜一搜看一看。搜索表达式：[“号陕”]. 真是让人大开眼界：在第 1 页的前 10 个结果中，不仅都不是词，而且除了第一个结果两个字是连在一起的外，其余 9 个两字之间都有或多或少的标点符号（包括空格）。再看后面的若干页，基本上也是如此。这样就不难理解了：为什么感觉中根本不是词甚至很少在一起用的两个字，会排在候选的首位。见表 1 和图 2(a).

但这样的搜索结果有什么用呢？本来是想搜索连续字串，结果出来的却多是中间插入了标点符号。如果想看一下“号陕”两个字紧密相连、中间未被其它符号隔开的情况有多少该怎么办呢？没有办法。在直觉上，“号陕”的搜索结果就应该是两个字“亲密无间”的；但事实上，却是大多都被隔开的；更大的问题是，直觉上合理的结果没有可以实现的途径。

是不是对“号陕”这样很少见的组合的搜索没有什么意义？或者这只是在汉语中才存在的问题？下面用一个英语的例子来说明。

表 1 “号陕”的前 10 个结果概要中的关键词与上下文

（关键词边框为本文作者所加，以便于阅读）

关键词与上下文	内嵌标点符号
陕政〔2008〕31 号陕县人民政府	
广告批准文号：陕药广审（文）第 2007010008 号	：
【法律文号】：陕劳社发[2004]123 号	】：
www.tczyxy.net 备案号：陕 ICP 备 06008556 号	：
陕 K 三号- 陕 K 三号- 和讯个人门户	-
许可证号：（陕）字第 2001213#	：（
国庆节专号------陕北行-	------
文 号：陕政发[2007]26 号	：
陕价费调发[2001]29 号、陕价费调发[2003]3 号	、
秦单 4 号 陕单 16 陕单 21	[空格]
秦椒 2 号 陕椒 2001 陕彩椒 1 号陕彩椒 2 号陕椒 2006	[空格]

3.3 英语中的问题同样存在

在 1.2 中说过，刚刚开始萌芽的思想往往都是出现次数不多的，并且很可能暂时被已有的声音淹没。上面的“号陕”就是一例，紧密组合的情况淹没在内嵌标点的例子中。

作者在实际的研究工作中遇到的一个实例，希望查询”similarity entropy”的出现情况。搜索表达式：[”similarity entropy”]. 查到的结果据 Google 称有上百个(实际打开有 50 多个，加上重复的共 80 多个)，见图 2(b). 但绝大部分都是被标点（逗号、句号、分号等）分开或连字号连上的。经检查，中间有连字号的多为并列关系，被其它标点分开的基本上没有直接关系。而实际希望的结果是中间为空格的，即为修饰关系的，经检查，这样的结果只有 3 个，如表 2 所示。

3 个真正相关的结果，混杂在 50 多个结果中，而且都不靠前，这显然增加了筛选的成本和负担。如果说，形式上是完全一样的，没有办法进行区分，只有人去甄别，那也罢了。问题是，形式上完全可以区分的，却因为 Google 的视同无别，才导致混淆不清。Google 为“确解用户之意”做了很多涉及语义消歧的努力，但如果不先去解决简单地在形式上加以区分就能解决的问题，恐怕要想“切返用户之需”就得付出更久的努力。

表 2 **“similarity entropy”**的 50 多个结果中词间为空格者（只有 3 个，且都不靠前）
（每项内容依次为 标题 1 行、概要 2 行、URL 1 行）

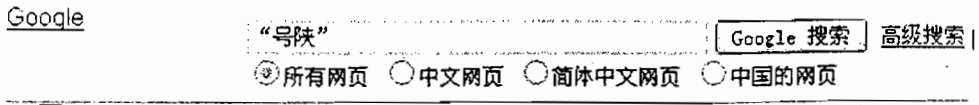
Developing higher-order networks with empirically selected units ... We shall introduce first the notion of **similarity entropy**, ... Thus the computation of **similarity entropy** in the Algorithm ... ieeexplore.ieee.org/iel4/72/7656/00317722.pdf
熵能的查询结果 --cnki 翻译助手 The concepts of the similarity element andsimilarity system and **similarity entropy** are proposed. The paper goes furtherto explore the relationship between ... dict.cnki.net/dict_result.aspx?searchword=熵能
Сервер періодичних видань::Каталог статей Three new concepts, i. e. , Similarity Unit, Similar Systems and **Similarity Entropy** are proposed. Two different kinds of distinction of the Similarity Unit ... scilib.univ.kiev.ua/article.php?658768

4 外部链接视同文档内容？——“codefusion” 与 “coldfusion”

在 Google 搜索中，如果用“网页快照”来打开，有时候会发现这样一个现象：输入的搜索词并没有在网页内容中出现，Google 告诉你“这些搜索字词仅在指向此网页的链接中出现”。这就让人很疑惑：既然网页内容中没有出现，为什么还要把它列出来？先看一个例子，再听 Google 的解释，然后做浅显的分析。

4.1 问题的引出

近来 U 盘病毒相当猖獗，作者在分析 U 盘病毒的传播机制时，发现 AutoRun.inf 是一个关键的环节，如果能把它变成一个个性化的名字，就能够防止病毒借其自动运行。为此查找相关工具，发现一个名叫 CodeFusion 的软件。搜索表达式 [CodeFusion 使用] 查出来的几乎都是关于 ColdFusion 的(前 10 个中的前 9 个)，搜索表达式 [CodeFusion 是什么] 的结果更是如此，前 10 个都是关于 ColdFusion 的，甚至前 100 个中都没有关于 CodeFusion 的，见图 3(a). 搜索表达式 [what’s codefusion] 也是相差无几，见图 3(b). 值得注意的是该图中的提示信息“您是不是要找：what's ***coldfusion***”，这说明 Google 很清楚用户现在查的不是 coldfusion, 而且要返回关于 coldfusion 的结果也应该在用户点击确认之后。从这两个例子可以看到，不管中文英文，字符集的大小，都有这个问题。

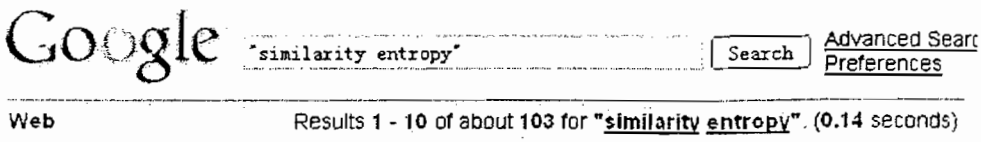


长安路181号陕教社家属院（陕师大南）-赶集网房产
长安路181号陕教社家属院（陕师大南）. 广告代码：029020141746 发布时间：2007-12-12 17:15:43. 价格：1000元/月 户型：2居 使用面积：65平方米. 地段：碑林-长安路- ...
xa.ganji.com/housing1d/07121217_41746.htm - 11k - 网页快照 - 类似网页

油菜新品种"陕油6号"
油菜新品种"陕油6号". 陕油6号油菜是西北农林科技大学育成的新品种。2000年元月通过陕西省农作物品种审定委员会审定。该品种具有以下特点： 1、 丰产性好 ...
kych.nwsuaf.edu.cn:5030/content/outcome/sy6.htm - 13k - 网页快照 - 类似网页

转发陕教明电〔2006〕12号、陕教稳〔2006〕21号、公传发〔2006〕1477号 ...
渭滨区教育局转发陕教明电〔2006〕12号、陕教稳〔2006〕21号、公传发〔2006〕1477号文件的通知. 各镇（乡）教育组、中小学、幼儿园： 为了有效地防止师生伤害事故的 ...
www.wbjy.net/newsInfo.aspx?pkId=3613 - 56k - 网页快照 - 类似网页

a) “号陕”



Static Symmetry and Dynamic Symmetry at Critical Point
After rejection of the Gibbs paradox statement (discontinuous **similarity-entropy** relation) and higher symmetry-lower entropy in statistical mechanics ...
flux.aps.org/meetings/YR00/MAR00/abs/S7150065.html - 3k - Cached - Similar pages

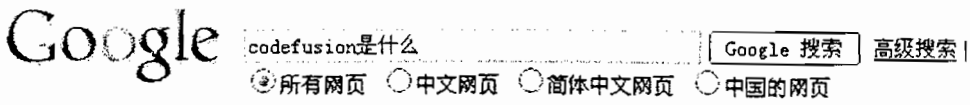
Entropy-based vs. similarity-influenced: Attribute selection ...
Case based reasoning ; Statistical analysis ; Electronic trade ; Information measure ; Information use ; **Similarity ; Entropy** ; ...
cat.inist.fr/?aModele=afficheN&cpsidt=14841061 - Similar pages

Image matching using alpha-entropy measures and entropic graphs ...
Similarity ; Entropy ; Pattern matching ; Mutual information ; Feature extraction ; Image matching ; Minimal spanning tree ; Image registration ; Image ...
cat.inist.fr/?aModele=afficheN&cpsidt=16383304 - Similar pages

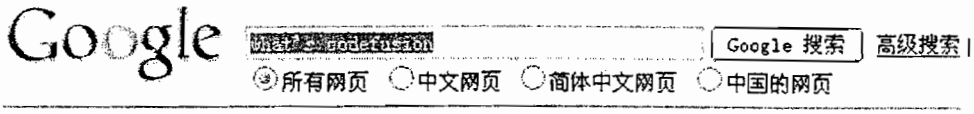
[PDF] Diversity Assessment Based on a Higher Similarity-Higher Entropy ...
File Format: PDF/Adobe Acrobat - View as HTML
already illustrated our unique approach based on our **similarity-entropy** ..... **similarity. Entropy** decreases discontinuously with the similarits of the ...
arxiv.org/pdf/physics/9910032 - Similar pages

b) “similarity entropy”

图 2　搜索连续字符串，标点符号当中嵌



a) codefusion 是什么



b) what's codefusion

图 3 搜索何为 codefusion，结果却是 coldfusion

那会不会是因为有人用 CodeFusion 制作破解补丁而被列为不受欢迎的软件呢？CodeFusion 只是一个补丁制作工具，完全不具备恶意软件的特征，没有任何理由封杀。就像菜刀在罪犯手里也可能成为凶器，并不能因此就不让人们使用菜刀。事实上，单独搜索[codefusion] 出来的都是和 CodeFusion 相关的，而且数以万计，这也说明 Google 并没有屏蔽 codefusion，因此前面的结果就更能凸显出 Google 语言模型的问题所在。

4.2 Google的解释

在 Google 提供的帮助中，“我的搜索结果：搜索字词不在网页上”[7]是这样解释的（仿宋体为原文引用）：

> 有时，Google 会将不包含您搜索的文字或词组的网页列入您的搜索结果。即使进行词组搜索也可能会出现这种情况。在评估网页的价值和相关性时，Google 不仅会考虑网页本身，还会考虑指向此页的链接的定位文字。如果指向网页的链接包含您搜索的词组，Google 可能会将此网页作为符合查询的匹配项返回。如果发生这种情况，网页的网页快照会显示"这些字词仅在指向此页的链接中显示"。
>
> 如果您关注的问题是搜索结果中的某网页并未包含您搜索的短语，建议您与链接该网页的网站管理员联系。...
>
> 如果链接网站的管理员受理了您的请求，下次抓取后我们的搜索结果会反映这些更改。

这说明 Google 认为其这样做是合理的，是为了更准确更有效地评估网页的价值和相关性。但我们看到的情况好像不是这样。正是由于 Google 统一用算法来排序而不进行单独调整，一旦出现典型性的反例，才更能说明其语言模型中的带有普遍性的问题。

4.3 浅显的分析

我们在此做一个浅显的分析：外部链接中的标识文字能否不加区别地与网页内容中的文字等同看待，作为搜索字词是否出现的硬性指标？

在 Shannon(1948)的通信模型中，更多的关心的是编码/解码的技术层面，是关于信息的形式（语形）的，committer 和 receiver 可以理解为“发报人/发报机”和“收报机/收报人”构成的“人 和/或 机”的人机系统，完成“消息”和“信号”之间的转换工作（编码/解码），凡是和语义相关的部分都是由其中的“人”来完成的。

在 Jakobson(1960)的通信模型中的 sender 和 receiver 已经是“发送者”和“接收者”，形成了“作者”和“读者”模式。谢清俊、谢瀛春则更明确地把传播过程中的“创作端”和“接收端”作为定义信息的两个基本立场，“作者”情境和“读者”情境的划分非常清晰。

在创作端，作品内容是由作者构造的，作者对作品内容具有控制权；在接收端，作品内容是由读者解析的，读者对作品内容必须有知情权。作者可控、读者可知，这是对传播中的信息的最基本要求。返观指向网页的链接文字，在通常情况下，不经非常的额外努力，是作者不可控的、读者不可知的，因而是和这一非常浅显的原则相背离的。4.2 节的引文中，第 2，3 段是 Google 提供的解决方案，显然是代价非常高的，而且如果链接该网页的网站管理员不接受修改请求的话，这个问题是无法解决的。

链接信息并非不可用，但它只宜作为软性指标，用来调整(0, 1)间的相关度，而不宜用作硬性指标来判断{0, 1}相关性。让上帝的归上帝，凯撒的归凯撒。

5 结论和展望

本文通过典型性实例讨论了 Google 语言模型的两个并非偶然性导致的缺陷：内嵌标点也算连续字串，外部链接视同文档内容。前者导致“号陕”超过“好山”，后者导致查[code…]出[cold…].

本文完全从学术的角度出发，不涉及任何与政治有关的问题。也就是说，本文所指出的缺陷，完全是 Google 本身的缺陷，而不仅仅是谷歌(中国)的问题。

本文 3，4 部分指出的最关键的两点缺陷，Google 如果能予以弥补，当然是我们希望看到的；如果 Google 坚持以前的做法，不认为这是缺陷（或者正是自己的特色），用户也会有自己的判断。

参考文献

[1] The American Customer Satisfaction Index, http://theacsi.org/

[2] Shannon, C. A Mathematical Theory of Communication. Bell System Technical Journal, 27(1948), 379--423, 623--656.

[3] Zipf, G. Psycho-Biology of Languages. Houghton-Mifflin, 1935.

[4] Anderson, Chris. The Long Tail: Why the Future of Business Is Selling Less of More. New York: Hyperion. 2006.

[5] Google, 1T Web 5-gram, (LDC2006T13), http://www.ldc.upenn.edu/

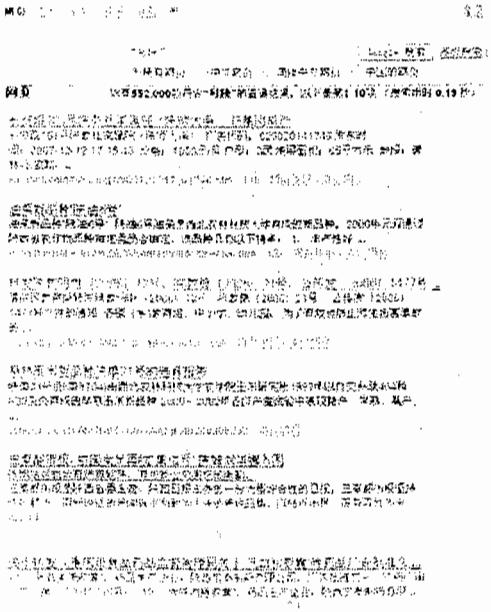
[6] Google, 谷歌拼音输入法, http://tools.google.com/pinyin/

[7] Google, 我的搜索结果：搜索字词不在网页上. (My search results: Search terms not on page.) http://www.google.cn/support/bin/answer.py?answer=427

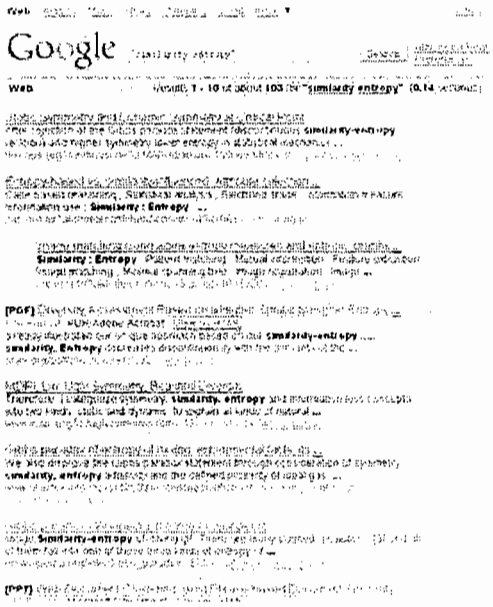
[8] Jakobson, R. Linguistics and Poetics, in *Style in Language*, MIT Press, 1960, 350-377.

[9] 谢清俊，谢瀛春. 一个通用的资讯（信息）定义. (A General Definition of Information.) http://pnclink.org:8080/pnc2006/A_General_Definition_of_Information.pdf pnclink.org:8080/pnc2006/Presentation%20material/keynote%20speech%20--C.C.%20Hsieh.pdf

附录 部分大图的缩图（左边有汉语，右边仅英语；上面不准确，下面不符合）

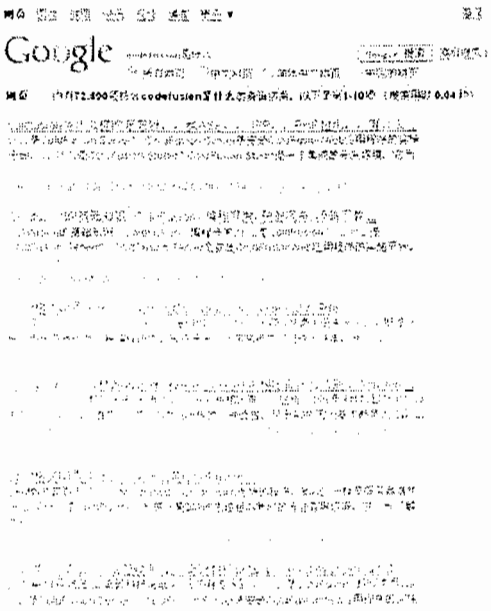


a）搜索："号陕"
（前 6 个中有 5 个 字间有标点符号）

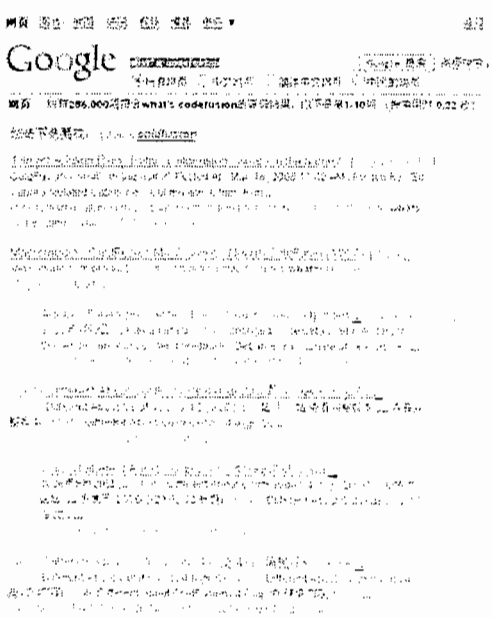


b）搜索："similarity entropy"
（前 7 个 词间都不是空格）

图 2 搜索连续字符串，标点符号当中嵌



a）搜索：codefusion 是什么
（前 6 个都是 coldfusion）



b）搜索：what's codefusion
（前 6 个都是 coldfusion）

图 3 搜索何为 codefusion，结果却是 coldfusion