基于超链接的镜像页面比较策略研究*)

杨 楠
(中国人民大学信息学院 北京 100872)

摘 要 Web中存在大量的镜像页面,这会严重影响分析的结果,并且占据大量的空间和资源,严重影响了计算的效率,因此,如何去除这些镜像页面是社区发现技术中的一个重要的问题。对基于纯链接的镜像页面去除方法[1,2]进行了分析,并证明了只需出度邻近的页面进行比较,并提出了页面邻近区域的比较方法,按照Web页面的分布理论,设计了比较策略的方法。实验结果证明,大大减少了比较的次数,提高了效率。

关键词 链接分析,镜像页面,页面相似度

Research on Comarison Strategy of Mirror Pages Based on Hyper-Links

YANG Nan
(School of Information, Renmin University of China, Beijing 100872)

Abstract There are many duplicated pages in Web. These mirrors of pages will distort the analysis result. The duplicates also occupy much space and resources, degrading system efficiency. How to delete these duplicates is a very important issue. The thesis analyzes the deleting method of duplicated pages based on hyper-links and proves that only neighboring comparison is required. The neighbor comparing method is proposed according the Web distribution on out-degree. The result of experiment shows that the comparing amount has been cut down dramatically and the computing efficiency is improved.

Keywords Link analysis, Duplicated pages, Page resemblance

1 引言

Web中存在着大量的镜像或重复页面,例如,大量的Yahoo! 的镜像页面,而其中许多页面并不属于Yahoo的。主要是由于站点镜像、相同页面的别名所造成的,另一个重要原因是Web页面的内容复制起来非常容易,因此,大量的页面只有很少的不同。许多页面的产生仅仅是少许的修改(例如,拷贝内容之后,修改作者的信息)[3]。统计表明,镜像网页数占总网页数的比例高达22%[4]。

镜像页面的存在会导致结果的严重失真,会形成一些假的社区,严重干扰了正常的社区抽取过程。同时,大量的镜像页面会占据大量的系统资源,降低了算法的效率。

因此,如何在应用社区发现算法之前删除这些镜像页面是非常重要的。我们对以前社区发现算法中所采用的去除镜像页面方法进行了分析。由于数据集很大,两两页面比较的方法不太现实,因而需要采用分组比较的方法。而以前的方法没有涉及到如何分组,以及分组的策略等问题。我们根据Web页面的出度分布,从理论上证明了无需所有页面之间的两两比较,而采用区域比较的方法,并提出了页面数据集分组的策略和计算方法。该方法可以大大减少比较的次数。

第2部分介绍了相关的研究内容;第3部分介绍了采用链接判定重复页面所采用的策略;第4部分是实验数据和结果分析;最后是结论和未来的工作。

2 相关研究

2.1 页面的相似度判定

文档的相似度判定方法是为每个文档计算出一组指纹(fingerprint),若两个文档拥有一定数量的相同指纹,则认为这两个文档的内容重叠性较高,也即二者是近似镜像的。

假设有文档A、B,$S(A)$和$S(B)$分别表示A和B的指纹集合,因此,文档A和B的相似度r可以定义为:

$$r(A,B)=\frac{|S(A)\cap S(B)|}{|S(A)\cup S(B)|}$$

其中,$|A|$表示集合A的大小。相似度r是介于0和1之间的数,并且$r(A,A)=1$。

社区发现技术中,由于页面的数量很大,因此基本都采用纯链接的方法,而避免对页面文档内容的处理,来减少计算代价。所以,指纹的产生是通过对每个页面中抽取的链接计算得到的。

2.2 采用语义聚类方法(Syntactic Clustering)[4]

文[1,2]和文[5]中均采用语义聚类方法删除镜像或近似的页面。以下是对语义聚类方法的介绍。

首先,需要文本的相似度定义。将每个文本抽象为词(Word)的序列,对这些序列进行词汇的分析使之成为规范的标记序列,而这一序列不理会HTML文本的细节,如格式、HTML命令、大写等等,然后将每个文本D和标记$S(D,w)$的序列集合关联起来。

包含在文本D中的一个连续序列我们称之为片段(Shingle)。对于每个文本D我们都可以定义它的w片(w-Shingle),用$S(D,w)$表示。这样,$S(D,w)$表示文本D包含的w个唯一表示的片的集合。例如,文本(a, rose, is, a, rose, is, a, rose)的4片段就是以下的集合:

*)本文得到教育部211项目子课题《WEB资源发现技术研究》的资助。杨 楠 博士,副教授。

{(a,rose,is,a),(rose,is,a,rose),(is,a,rose,is)}

对于一个给定大小的片段尺寸，我们可以得到文档 A 和 B 若干个片段。每个片段用一个指纹函数来表示，这样文档 A 和 B 可以抽象为两组指纹函数，所以，文档 A 和 B 的相似度 r 可以用前面的公式计算。

Web 社区发现算法中，拖网算法[1,2]的二分核算法均采用了这个方法去除镜像或近似页面。由于针对的是海量的数据级别，因此需要对片段(Single)的摘取仅针对页面内部的链接序列，而不是整个 HTML 文档。该方法仅使用了 Web 页面链接的局部哈希部分，并且比较了很少的几个哈希值(即片段)来检测重复页面。Broder[3]证明，如果哈希函数和片段的数量选择得当，可以有很高的概率去除相同或者几乎相同的重复页面。

文[1]选择的策略是每个页面仅仅对应 2 个片段，每个片段中有 5 个链接。文[2]选择的策略是每个页面对应 3 个片段，每个片段内包含 4 个链接。文[1]实验结果证明，最后人工检查了被删除的页面，基本上都是镜像页面。

2.3 采用直接统计的方法

另外一种较为直观的方法就是将页面的所有链接变换为对应的指纹，这样所有的页面都被抽象为一组指纹的集合。直接利用 2.1 节的方法就可以判定 2 个页面的相似度。例如，文[7]中关于重复页面的定义如下：近似重复(near-duplicates)的两个页面应该满足 2 个条件：(a)每个页面的出度大于 10；(b)它们的链接至少有 95%相同。文[5]中关于重复页面和文[7]相同：(a)每个页面的出度至少大于 8；(b)它们之间的链接至少有 80%相同。

和 2.2 节的方法相比，这种方法的优点在于不需要选择哈希函数和片段的计算，计算方法简单，实现容易，比较客观地反映了页面的真实情况，但处理量比 2.2 大。

但是，这两种方法所涉及的是小数据集，在面临大量数据集的情况下，需要认真研究比较的策略。

3 方法描述

这两个算法看起来都非常简单，但是，采用简单的两两比较的方式确是不可能的。如果存在 N 个页面，比较的次数将达到 $O(N^2)$。假如实验数据集中有 6,000,000 个页面，比较次数将达到 10^{13}，很明显计算量过大。

因此，在语义聚类方法的实现上，采用处理大量数据的简单方法是，分割-计算-合并。将大量的数据分割为小的块，对每块分别计算，然后合并结果。如果整个集合中文档的数量是 N，将 N 个文档分割为 k 个小组，每组内的文档数为 n_1，n_2，…，n_k，$i=1,2,\cdots,k$，则计算代价从原来的 $O(N^2)$ 下降为 $O(\sum_{i=1}^{k}(n_i^2))$。

3.1 页面的出度分布和比较的关系

Web 页面链接数量的分布不是均匀的，对 Web 每个页面分布的大量统计表明，页面的连接入度和连接出度符合 Power-Law 定律[6]，即具有 i 个链接入度页面的数量正比于 $i^{-2.1}$，具有 i 个链接出度的页面数量正比于 $i^{-2.7}$。页面的出度和数量之间的关系分布如图 1 所示。

从图 1 可以看出，当出度超过某个值的时候，页面的数量急剧下降。而数量最大页面是在出度为 1 到 10 之间。而出度太小的页面之间作相似比较，公共的出度数量不能说明页面的相似性。假如，两个出度为 5 的页面相比，仅 4 个出链接相同，就可以得到相似度为 4/(4+5−4)=0.8。因此，我们需要限定最小出度。

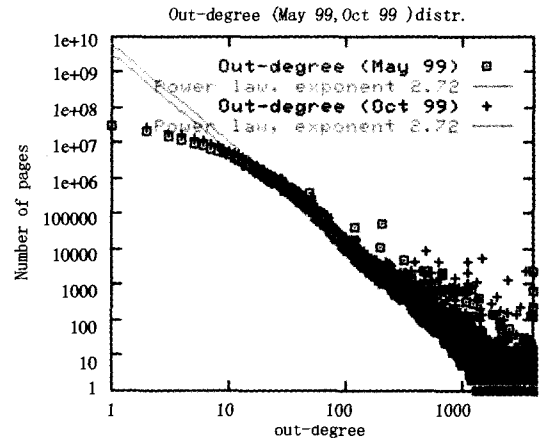

图 1 页面出度分布符合 Power-law 定理

我们设定最小的出度限制为 D_{min}，我们也采用文[7]和文[5]中的取法，让 $D_{min}=10$。这样，凡是出度小于 10 的页面都被忽略。另外从图 1 中可以看到，当出度大于某个值之后，页面的数量很小，而且基本上变化不太。因此，对于比较分组的可以考虑 2 个参数。页面的出度下限取值为 D_{min}，上限取值为 D_{max}。要取得 D_{min} 和 D_{max} 的值，要对得到页面集合的出度作 Power-Law 曲线分析，找到页面急剧变小的出度值作为 D_{max} 的值。

由于是针对 Web 页面数据集的境况，而且我们仅仅考虑每个页面的链接信息，因此，我们采用和语义聚类方法不同的比较策略。考虑到近似页面之间的链接出度差不应该太大。如果两个页面的出度差很大，就没有比较的意义。因此，我们将页面的两两比较限定在一定的区域范围之内。简单地说，就是仅比较出度相差不太大的页面。

假设，有 2 个页面 $p1$ 和 $p2$。页面 $p1$ 的出度为 n，页面 $p2$ 的出度为 m。又假设两个页面的相同的出链接数达到最大的可能值，即相同出度数为 $\mathrm{Min}(n,m)$。此时 2 个页面的重叠度最大，即 $S(A)$ 和 $S(B)$ 的交集达到最大，也就是 $|S(A)\cap S(B)|=\mathrm{Min}(|S(A)|,|S(B)|)=\mathrm{Min}(n,m)$。如果，$p1$ 和 $p2$ 出度小的表示为 $\mathrm{Min}(n,m)$，那么另一个页面的出度一定可以表示为 $\mathrm{Min}(n,m)+|n-m|$，因此两个页面的出度之和可以表示为 $\mathrm{Min}(n,m)+\mathrm{Min}(n,m)+|n-m|$。因而有 $|S(A)\cup S(B)|=|S(A)|+|S(B)|-|S(A)\cap S(B)|=2\mathrm{Min}(n,m)+|n-m|-\mathrm{Mim}(n,m)=|n-m|+\mathrm{Min}(n,m)$。

按照前面关于相似度的定义，则有：

$$r(A,B)=\frac{|S(A)\cap S(B)|}{|S(A)\cup S(B)|}=\frac{\mathrm{Min}(n,m)}{|n-m|+\mathrm{Min}(n,m)}$$

如果考虑到相似度大于 α，则上述公式变为：

$$r(A,B)=\frac{\mathrm{Min}(n,m)}{|n-m|+Min(n,m)}>\alpha$$

最后得不等式：

$$|n-m|<\mathrm{Min}(n,m)\frac{1-\alpha}{\alpha}$$

也就是说，只有当两个页面的出度之差小于 $\mathrm{Min}(n,m)\alpha$ 的时候，才可能使得相似度大于 α。因此，我们得出结论，就

是一个出度为 n 的页面，当和出度为 m 的页面进行形似度比较时，当 $|n-m|$ 大于 $\mathrm{Min}(n,m)\frac{1-a}{a}$ 的比较是没有意义的。

3.2 比较区域的确定和比较策略

根据前面的结论，不必要所有页面参加比较，只要比较出度在某个区域之内的页面。首先，介绍一下基本的比较策略。

如果存在一个有 n 个元素的数据集 $S=\langle d_1,d_2,\cdots,d_n\rangle$，并且如果 $\langle d_i,d_j\rangle$ 的比较和 $\langle d_j,d_i\rangle$ 的结果是等价的，那么，我们采用每个元素和它前面所有元素比较的方法。例如，数据 d_i 和 $d_1,d_2,\cdots,d_{i-1}$ 比较，d_1 的比较次数为 0，d_n 的比较次数为 $n-1$。所有的比较次数为 $n\times(n-1)/2$。

如果存在两个不相交的数据集 S_1 和 S_2，其中 S_1 有 n 个元素，$S_1=\langle x_1,x_2,\cdots,x_n\rangle$；$S_2$ 有个 m 元素，$S_2=\langle y_1,y_2,\cdots,y_m\rangle$。要实现这两个数据集之间所有元素的比较，比较次数为 $n\times m$。

因此，如果将整个页面的集合作为一个数据集来看待的话，如果整个页面的个数是 N，则需要的比较次数是 $N\times(N-1)/2$。

根据前面的结论，不必要所有的页面都参加比较。因此，我们可以将整个页面的集合划分为若干不相交的子集。选择合适的子集参加比较。这样就可以大大减少总的比较次数。

下面我们将设计比较的策略。

1）按照出度对页面排序

首先将所有的页面按照出度排序，并且将相同出度的页面分在同一组。如果整个集合页面的出度的范围是 $[0,k]$，就形成了 $k+1$ 个小组，每个小组内包含所有出度相同的页面。这样将整个页面分割成为 $k+1$ 个不相交的子集。

2）确定最小出度和最大出度值

由于出度较小的页面相似度的比较意义不大，例如文[5]对出度大于 8 和文[7]采用出度大于 10 的页面参加比较，我们采用文[7]的方法，只比较出度大于 10 页面。另外，从页面出度分布情况看，当页面的出度大于某个值的时候，页面的数量急剧下降，并且大于某个值页面的出度相差较大，数量不多，存在镜像或近似页面的概率较小。因此，我们采用当出度大于某个值时的所有页面不参加比较。我们限定最小的出度为 $D_{\min}$，最大的出度为 $D_{\max}$，这样页面的比较范围就是 $[D_{\min},D_{\max}]$。

3）页面的存储

每个页面都是由该页面的 URL 来表示，但是为了便于存储和表示其唯一性，我们将每个页面的 URL 都转换成指纹函数来表示。我们采用 128 位的 MD5 指纹函数，这样，每个页面均可表示成为一个 16 位字节的数。

4）比较策略设计

由于在任何一个子集当中，页面都是以指纹函数的形式表示，不可能存在两个相同的 URL，因此不可能存在两个相同的指纹函数。我们将指纹函数在子集内排序，所以，子集内的页面也是有序的排列。

因此，从全局的角度上看，任何一个页面 P 可以表示为 $P_{i,j}$，其中 i 是所属子集 $S_{i,j}$ 表示在子集 S_i 内指纹函数排序的第 j 个。按照前面的基本比较策略，我们对于任何一个页面，只比较排在前面的所有页面。

因此，对于任何一个页面，只需要和子集内排在自己前面的页面比较，还需要和排在该子集前面的若干子集中的页面比较。例如，对于任何一个页面 $P_{i,j}$，子集 S_i 内的比较为对 $P_{i,1},P_{i,2},\cdots,P_{i,j-1}$ 页面的比较。另外，假设比较区域为 δ，还需要对 $i-\delta,i-\delta+1,\cdots,i-1$ 子集内所有页面的比较。

下面，我们介绍 δ 的计算方法。按照 3.1 节的定义，要使得相似度大于 a，就要使得下列不等式成立：

$$|n-m|<\mathrm{Min}(n,m)\frac{1-a}{a}$$

由于比较的方向是和排列在前的页面比较，我们取当前页面的出度为 m，被比较页面的出度为 n，因此，$m>n$ 总是成立的。所以，$\mathrm{Min}(n,m)=n$。$|n-m|=\delta$，所以取 $\delta=\lceil n\frac{1-a}{a}\rceil$ 就可以满足上述不等式。

5）防止重复比较策略

由于比较过程是按照子集顺序、子集内顺序扫描的每个页面 $P_{i,j}$，因此，有些页面可能会在随后的比较时已经是被删除的页面。为了防止重复比较，我们为每个页面保留一个删除标记，当一个页面确定被删除时，将删除标记置 1，随后的比较如果发现被比较页面的删除标记为 1，则放弃比较。

4 实验数据和结果

4.1 数据收集

我们在 Web 社区发现技术的研究过程中，需要镜像页面的删除过程。实验中用到了一个数量为 270000 页面的数据集。该数据集所有页面的出度从 0 到 1005。其分布如下：

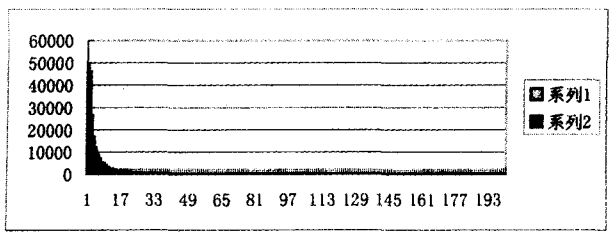

图 2 数据集中页面出度的分布

我们根据出度的分布情况，选择最小出度值 $D_{\min}=10$，最大出度值 $D_{\max}=150$。

4.2 实验结果

每个子集 S_i 比较范围的计算结果见表 1。

从表中我们得到总比较次数 84592370 次，这要比完全比较次数(270000)×(270000)/2=36450000000 减少了很多。

结论 由于大量的 Web 页面存在，我们通过基于纯链接下镜像或近似页面的去除方法的比较策略的研究，得出不必所有的页面两两比较，从页面的出度考虑，可以减少一些比较的次数，大大减轻了系统的负担。

但是，我们可以看到，当页面数量较大时，我们依然面临着巨大的比较数量。因此，我们将考虑进一步减少比较次数的方法。例如，考虑如何选择片段(shingle)的方法来减少比较的次数。我们未来的工作将进一步考虑和文档内容结合起来判断，而提高去除镜像页面的准确性。

表 1 比较次数的统计表

数据集组比较	比较范围 δ	比较次数
(10,10)比较：	1	(3171 * 3170)/2 次
(11,10)比较：	1	2748 * 3171 次
(11,11)比较：		(2748 * 2747)/2 次
(12,11)比较：	1	2445 * 2748 次
(12,12)比较：		(2445 * 2444)/2 次
…	…	…
(150,136)比较：	16	9 * 12
(150,137)比较：		9 * 9
(150,138)比较：		9 * 16
(150,139)比较：		9 * 22
(150,140)比较：		9 * 12
(150,141)比较：		9 * 9
(150,142)比较：		9 * 21
(150,143)比较：		9 * 10
(150,144)比较：		9 * 9
(150,145)比较：		9 * 10
(150,146)比较：		9 * 11
(150,147)比较：		9 * 17
(150,148)比较：		9 * 15
(150,149)比较：		9 * 10
(150,150 比较：		(9 * 8)/2
总比较次数		84592370

说明：表中(i,j)表示出度为 i 的子集和出度为 j 的子集的比较。

参 考 文 献

1 Gibson D, Kleinberg J, Raghavan P. Inferring Web Communities from Link Topology. In: Proc. of the 9th ACM Conf. on Hypertext and Hypermedia. Pittsburgh, PA, USA, 1998. 225～234

2 Reddy P K, Kitsuregawa M. Inferring Web Communities through relaxed-cocitation and power-law. KITSUREGAWA Lab: [Annual Report]. 2001. 27～46

3 Broder A, Galssman S, et al. Sytactic clustering of the Web. In: Proc. of the sixth international WWW Conf. April 1997. 391～404

4 Shivakumar N, et al. Finding near-replicas of documents on the Web. WebDB, 1998. 204～212

5 Wang Y, Kitsuregawa M. Clustering of Web Search Results with Link Analysis. KITSUREGAWA Lab: [Annual Report]. 2001. 18～26

6 Broder A Z, Kumar R, Maghoul F, et al. Graph structure in the Web. Computer Networks, 2000. 309～320

7 Dean J, Henzinger M R. Finding Related Pages in the World Wide Web. In: Proc. of the eighth international WWW Conf. Toronto, Canada: 1999. 389～401

(上接第 115 页)

三级网络构成一个串联的网络，串联后的网络提供的总服务曲线 β 为三个子系统提供的服务曲线 β_i，$i=1,2,3$ 的最小加卷积，即 $\beta=\beta_1\otimes\beta_2\otimes\beta_3$。如果三个网络均采用延迟—速率服务曲线，则串联后总的服务曲线为：$\beta=\beta_{R1,V1}\otimes\beta_{R2,V2}\otimes\beta_{R3,V3}=\beta_{R1\wedge R2\wedge\beta3,V1+V2+V3}$。

5 应用实例分析

设语音系统采用 G.729A 编码，语音帧的格式化时间为 $T_f=30$ ms，预处理时延为 $l=5$ms，总的端到端传输时延、排队时延和传播时延分别为 5、80 和 30 ms。设语音允许的最大单向端到端延迟为 $d_{max}=250$ ms。

假定系统采用的编码规则为 G.729A，每 10ms 产生一个数据帧，每个数据帧长 10 字节。通常一个 IP 包包括 3 个数据帧，则单方向产生数据包的速率为 33 包/s，每个数据包包含 30 字节的语音编码数据。在以太网上传输时一个 IP 电话数据包的大小为 84 字节。取峰值速率为平均速率的 4 倍，则整形系统的参数为 (M,p,b,r)=(84 字节，11088 字节/秒，400 字节，2772 字节/秒)。采用 IP 网络，假定语音数据流的传输路径长度为 5 跳，各跳的 MTU 均为 1518 字节，链路速率为 $c=10$Mbps，网络节点采用 WFQ 调度策略。

根据上述给定条件，系统允许的最大回放时延为：

$D_{play}=250-(30+5+5+80+30)=100$ms

$C_{tot}=5*M=420$ 字节，$D_{tot}=5*MTU/c=6.072$ms。

$T=(b-M)/(p-r)=40$ms

$\bar{d}=\frac{M+C_{tot}}{p}+D_{tot}=51.5$ms，$\bar{d}<D_{play}$。

代入上面的式(5)，计算得：

$R=\frac{pT+M+C_{tot}}{D_{play}+T-D_{tot}}=7071$ 字节/秒$\approx 2.5r^{\circ}$。

$V_{tot}=\frac{C_{tot}}{R}+D_{tot}=65.4$ms

代入式(6)可以求得需要的回放缓冲区长度为：$B_{max}=b+rV_{tot}=581$ 字节。

结束语 本文深入研究了语音通信系统的特点和端到端时延的组成及其计算方法，提出一个简单可行的语音业务回放控制模型，采用双令牌桶对语音流量进行整形，充分考虑了语音通信的突发性特点。基于最新的网络演算理论，推导出了给定端到端时延、语音到达曲线和网络服务曲线条件下的语音回放点(回放时延)、需要分配的速率和需要的缓冲区长度的计算公式，并简单论述了由接入网和骨干网组成的多网络环境下上述公式的使用方法。最后通过应用实例分析验证了本文的分析结论。分析结果表明，对于 250ms 的单项端到端时延，100ms 左右的回放时延要求，需要分配的速率和缓冲区比较合理，带宽利用率达 40%左右。

参 考 文 献

1 Cruz R L. Quality of Service Guarantees in Virtual Circuit Switched Networks. IEEE Journal of Selected Area in Communication, 1995, 13(6): 1048～1056

2 Boudec J Y L, Thiran P. Network Calculus: A Theory of Deterministic Queuing System for the Internet. Heidelberg: Springer-Verlag, 2004

3 Firoiu V, Le Boudec J-Y, Towsley D, Zhang Zhi-Li. Theories and models for Internet quality of service. Proceedings of the IEEE, 2002, 90(9): 1565～1591

4 Yuming J. Relationship between guaranteed rate server and latency rate server. Computer Network s, 2003, 43(3): 307～315

5 Shenker S, Partridge C, Guérin R. Specification of guaranteed quality of service. RFC 2212, 1997

6 Wroclawski J. The use of RSVP with IETF integrated services. RFC2210, September 1997

7 Recommendation G.711, Pulse Code Modulation (PCM) of Voice Frequencies. ITU, Nov. 1988

8 Annex A to Recommendation G.729, Coding of Speech at 8kbit/s using Conjugate Structure Algebraic-Code-Excited Linear-Prediction (CSACELP), Annex A: "Reduced Complexity 8 kbit/s CS-ACELP Speech Codec". ITU, Nov. 1996

9 Recommendation G.723.1, Speech Coders: Dual Rate Speech Coder for Multimedia Communications Transmitting at 5.3 and 6.3 kbit/s. ITU, March 1996

10 Karam M, Tobagi F. Analysis of the delay and jitter of voice traffic over the Internet. In: Proc. of Infocom 2001. http://citeseer.ist.psu.edu/karam01analysis.html

11 Recommendation G.114. Transmission systems and media general characteristics of Internet telephone connections and Internet telephone circuits one-way transmission time. http://www.pesq.org, 1996

12 Gruber J, Strawczynski L. Subjective Effects of Variable Delay in Speech Clipping in Dynamically Managed Voice Systems. IEEE Transactions on Communications, 1985, COM-33(8)

13 Friedman T, Caceres R, Clark A. RTP control protocol extended reports (RTCP XR). RFC 3611. http://www.ietf.org, 2003