بسمه تعالی

فهرست مندرجات

۱- مقدمه

استراتژی های بازیابی میزان شباهت میان یک " پرس و جو "[1] و یک مدرک را بیان می کنند. اساس این استراتژیها بر اصل ارتباط بیشتر میان "پرس و جو" و مدارک استوار است . یک استراتژی بازیابی، الگوریتمی است که پرس و جوی Q و مجموعه ای از مدارک $D_1,D_2,..,D_n$ را گرفته و ضریب شباهت $SC(Q,D_i)$ را برای تمام مدارک محاسبه می نماید .

یکی از استراتژیهائی که در بازیابی اطلاعات تأثیرگذار می باشد استفاده از سیستم شبکه عصبی است. این استراتژی حاوی مجموعه ای از نرونها یا گره های شبکه است که به هنگام پرس و جو و در زمان بازیابی مدارک فعال می شوند . یک شبکه عصبی شامل گره ها و پیوندها می باشد که بر اساس مقادیر ورودی و خروجی سیستم تعریف می شود .

۱-۱ تعریف شبکه عصبی هوشمند

هم اینک از شبکه های عصبی هوشمند به عنوان مدلهای ارتباطی یا پردازنده های موازی توزیع شده یاد می شود. شبکه های عصبی به عنوان رده ای از مدلهای محاسباتی یا الگوریتمهایی که بر اساس تقلید از مغز انسان پیاده سازی شده اند مطرح اند. ایده طراحی شبکه عصبی از سیستم شبکه عصبی بیولوژی ناشی شده است. بعضی از ویژگی های اساسی عناصر پردازش شبکه عصبی که از خصوصیات نرونهای بیولوژی پیروی می کنند به شرح زیر می باشد:

۱. هر نرون دریافت کننده تعدادی سیگنال ورودی است .
۲. مقدار سیگنالهای ورودی ممکن است با وزنی که به آنها داده می شود تغییر نماید .
۳. نرون (پردازشگر) ورودیهای وزن دار را با یکدیگر جمع می کند.
٤. نرون ،تحت شرایط مناسب (ورودیهای کافی) ، سیگنال خروجی را منتشر می کند .

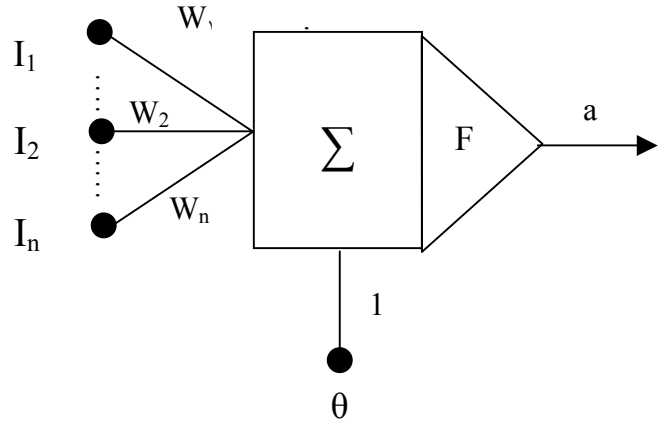
[1] -Query

۵. خروجی از یک نرون خاص ممکن است به چند نرون دیگر انتشار یابد.

تاریخچه استفاده از شبکه های عصبی هوشمند به سال ۱۹٤۰ باز می گردد و از آن زمان تاکنون پیشنهادهای بسیاری برای مدلهای محاسباتی شبکه عصبی ارائه شده است . آنچه در تمام این مدلها مشترک است وجود عناصر اولیه ای درتمام این مدلها است . به طوری که بیان شد ، یک شبکه عصبی هوشمند شامل تعداد زیادی از واحد پردازش ساده ای به عنوان نرون می باشد. برای ایجاد یک شبکه عصبی اتصال نرونها به یکدیگر با استفاده از پیوندهای وزن دار لازم است. نرونها با ارسال سیگنال به یکدیگر با هم در ارتباط می باشند .

۱-۱-۱ تعریف نرون

یک نرون واحد پردازشگری است که می تواند محاسبات ساده ای به انجام برساند . هر نرون ، ورودی را برای محاسبه از اتصالات ورودی دریافت کرده و نتیجه محاسبات به عنوان خروجی به نرون بعدی فرستاده می شود . شکل ۱ یک نرون را با n اتصال ورودی وزن دار نمایش می دهد :



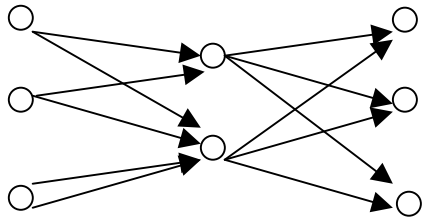
شکل ۱ – یک نرون با چند ارتباط ورودی

محاسبات با درگیری یک نرون با اولین ورودی شبکه شروع می شود . در اغلب شبکه های عصبی ، ورودیθ با وزن یک به عنوان بایاس وجود دارد . خروجی a نرون مطابق فرمول زیر محاسبه می شود:

$$a = F\left(\sum_{i=1}^{n} w_i I_i + \theta\right) \qquad (۱)$$

درحالیکه f تابع فعال سازی است به عنوان یک فاکتور کلیدی نیز که بیانگر رفتار نرون است شناخته می شود . در هر لایه معمولاً نرونها دارای یک تابع فعال سازی می باشند .

بسیاری از شبکه ها از تعدادی لایه که در هر لایه تعدادی نرون وجود دارد ، تشکیل شده است . شبکه های عصبی که شامل بیش از یک لایه از نرونها باشند شبکه عصبی چند لایه ای نامیده می شوند . شکل ۲ شبکه عصبی هوشمند را با سه لایه از نرونها نشان می دهد.



شکل ۲: شبکه عصبی هوشمند

بر اساس توپولوژی ساخت شبکه های عصبی ، این شبکه ها به دو نوع Feed forward و recurrent دسته بندی می شوند . درشبکه های چند لایه ای Feed forward معمولاً یک لایه ورودی وجود دارد که سیگنالهای خارجی را به عنوان ورودی شبکه قبول می کند. همچنین، لایه ای به عنوان لایه خروجی وجود دارد که نرونهای فعال این لایه به عنوان خروجی سیستم می باشد. ارتباط میان لایه ها به صورت Feed forward است . لایه های میان ورودی و خروجی به عنوان لایه های پنهان نامیده می شود. نمونه ای از شبکه های عصبی Feed forward ، پرسپترون[۱] و آدالین[۲] می باشد.

درشبکه هایrecurrent ، جهت اتصال نرونها از حلقه استفاده می شود. در این شبکه ها، هر سیگنال می تواند به تعدادی از نرونها متصل باشد . دراین حالت ، شبکه تا زمانی پردازش ها را تحمل می کند که مقادیر فعال ساز شبکه

[1] - Perceptron
[2] - Adalin

به حالت پایدار خود برسند. شبکه های هاپفیلد[1] نمونه ای از این توپولوژی می باشد .

۲-۱-۱ تنظیم وزن

به غیر از معماری شبکه ، روش تنظیم مقادیر وزنها (آموزش) یکی از مهمترین ویژگیهای تشخیص شبکه های عصبی مختلف می باشد . دو نوع آموزش وجود دارد :

۱. آموزش تعلیم یافته (Supervised)
۲. آموزش غیر تعلیم یافته (Unsupervised)

در روش آموزش یادگیری با معلم ، آموزش با نمایش دنباله ای از بردارهای آموزشی که هر کدام به بردار خروجی خاصی اختصاص می یابند آغاز می گردد . وزنها بر اساس الگوریتمهای یادگیری تنظیم می شوند . این پردازش به عنوان آموزش تعلیم یافته نامیده می شود . بعضی از شبکه های عصبی که برای دسته بندی الگوها به کار برده می شوند از این روش برای دسته بندی بردارهای ورودی خود به مجموعه های مختلف استفاده می نمایند .

در شبکه های عصبی غیر تعلیم یافته دنباله ای از بردارهای ورودی بدون مشخص شدن بردارهای خروجی تشکیل می شوند . وزنهای شبکه آن قدر تغییر می یابند تا بردارهای ورودی مشابه به یک واحد خروجی یا خوشه[2] اختصاص یابند . شبکه های عصبی[3] SOM از این دسته محسوب می شوند . از آنجایی که شبکه کوهنن[4] SOM در این طرح مورد استفاده قرار گرفته است به معرفی اجمالی این شبکه پرداخته می شود :

۳-۱-۱ شبکه SOM

کوهنن از سال ۱۹۸۸ به بعد ، مطالعات عمیقی را در باره شبکه ساده SOM به انجام رسانیده و در نهایت موفق به طراحی شبکه پیچیده SOFM

---

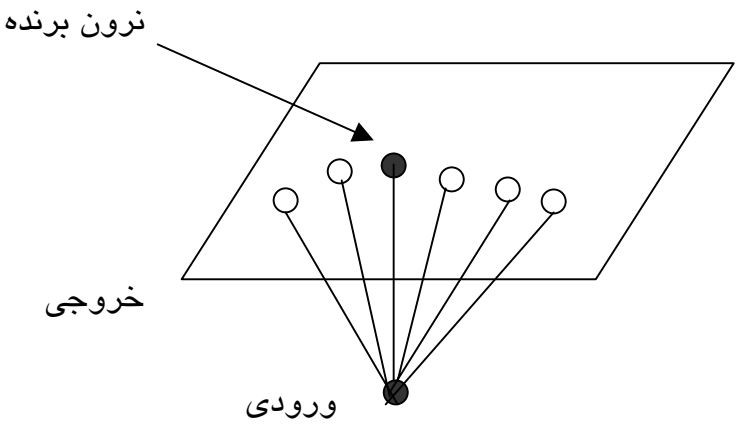
[1] - Hopfield
[2] -Cluster
3-Self - Organized Map
[4] -Kohonen

شده است . ویژگی شبکه SOFM سازماندهی آن است که از توانایی انتخاب همسایگی برنده به جای یک گره برنده برخودار است.

الگوریتم LVQ[1] که توسط کوهنن معرفی گردید جهت خوشه بندی بردارها در فشرده سازی داده ها به کار برده می شود . معماری شبکه کوهنن شامل دو لایه است :

۱. لایه ورودی
۲. لایه کوهنن (لایه خروجی)

این دو لایه کاملاً به یکدیگر متصل می باشد . هر نرون لایه ورودی، اتصالِ feed forward به تمام نرونهای لایه خروجی دارد . (شکل ۳)



شکل ۳ – مدل کوهنن (۱۹۹۹)

شبکه کوهنن در دو مرحله کار می کند :

۱. شبکه، واحدهایی را که بردار وزن ارتباطی آنها نزدیکی بیشتری به بردار ورودی جاری داشته باشد به عنوان واحد برنده انتخاب می کند .
۲. پس از انتخاب همسایگی برنده ، بردارهای اتصال به واحدهایی که مقدار خروجی آنها مثبت است به طرف بردار ورودی به چرخش در می آیند .

[1] - Learning Vector Quantization

ورودی به لایه کوهنن یا لایه خروجی می تواند با ضرب داخلی میان بردار وزن نرون و بردار ورودی محاسبه گردد . نرون لایه خروجی برنده ، نرونی است که بزرگترین ضرب داخلی را داشته باشد . زاویه میان بردار وزنی نرون برنده و بردار ورودی کوچکتر از زاویه با سایر نرونها است .

روش دوم جهت انتخاب نرون برنده انتخاب نرونی است که بردار وزنی آن دارای کوچکترین نرم فاصله اقلیدسی را از بردار ورودی داشته باشد . در این پروژه از این روش جهت انتخاب همسایگی برنده استفاده شد . الگوریتم زیر جهت آموزش شبکه SOM به کار برده می شود :

step0 initialize weights w_{ij}
set topological neighborhood parameters
set learning rate parameter
step 1 while stopping condition is false do step 2-8
step2 for each input vector x , do step 3-5
step3 for each j, compute:
$$d(j)=\sum_i (w_{ij}-x_i)^2$$
step4 find index J such that d(J) is a minimum
step 5 for all units j whitin a specified neighborhood of J
and for all i:
$$w_{ij}(new)=w_{ij}(old)+\alpha[x_i-w_{ij}(old)]$$
step6 update learning rate
step7 reduce radius of topological neighborhood at specified time
step8 test stopping condition

شکل ٤- الگوریتم آموزش شبکه SOM

۲-۱ بازیابی اطلاعات

سیستمهای بازیابی اطلاعات به جستجو و بازیابی داده ها از مجموعه ای از اسناد و مدارک در پاسخ به درخواست کاربر می پردازند .

از چند دهه قبل ، با ارزان شدن حافظه های کامپیوتری ساخت سیستمهای اطلاع رسانی الکترونیکی رو به افزایش نهاد . پایگاههای اطلاعاتی متن ، گرافیکی ، صوتی ، تصویری با حجم زیادی از اطلاعات ساخته شد .

با رشد روز افزون مجموعه سازی اطلاعات، نیاز به ابزاری دقیق تر برای بازیابی اطلاعات در پایگاههای اطلاعاتی حجیم هر روز بیشتر احساس می شود . اف. ویلفرید لانکاستر (۱۳۷۹) در کتاب خود می گوید نظام بازیابی اطلاعات، پدیدهٔ پیچیده ای است که شامل مدارک، تقاضاها (درخواستها) ، شرح مختصر این مدارک و تقاضاها، مکانیسم تطبیق آنها، و افراد می باشد. یک نظام بازیابی اطلاعات ممکن است متون کامل ، گزیده هایی از مدارک (مثل چکیده ها) ، یا نام و نشانیهای مدارک (یعنی ، اسنادهای کامل کتابشناختی) را بازیابی کند . نظامی که در نهایت متون کامل را برای استفاده کننده ، تهیه می کند نظام بازیابی مدارک نامیده می شود. یک نظام بازیابی به طور معمول در چندین مرحله عمل می کند (برای مثال ، نخستین برون داد آن ممکن است به شکل استنادهایی باشد که متقاضی از آنها می تواند دست به گزینش زند). پس از آن ، کاوشگر می تواند متون کامل مواردی را که برگزیده است در خواست نماید. ا. استون پولیت (۱۳۸۰) در کتاب خود معیارهای بازیافت و دقت بازیافت را برای توصیف عملکرد نظام بازیابی مورد استفاده قرار می دهد. این معیارها به شرح زیر تعریف می شوند :

کل تعداد مدارک مربوط ÷ تعداد مدارک مربوط بازیابی شده = بازیافت

کل تعداد مدارک بازیابی شده ÷ تعداد مدارک مربوط بازیابی شده = دقت بازیافت

این معیارها به وسیله روشهای فوق در واژگان کنترل شده تحت تأثیر قرار می گیرند :

ویژگی بالا – بازیافت کم

دقت بازیافت بالا

جامعیت بالا – بازیافت بالا

دقت بازیافت کم

ویژگی کم - بازیافت بالا

دقت بازیافت کم

جامعیت کم – بازیافت کم

دقت بازیافت بالا

تصور می شود که بعد از یک دوره زمانی که صرف جستجو می شود ، عملکرد کلی کاهش پیدا می کند، یعنی بازیافت در برابر کاهش شدید دقت بازیافت فقط به مقدار کم افزایش می یابد .

با افزایش حجم داده ها ، استفاده از روشهای سنتی بازیابی اطلاعات کافی نمی باشد . اکنون ، بازیابی اطلاعات با تحلیل ، نمایش و بازیابی متن سر و کار دارد . سیستمهایی که می توانند مدارک متنی را بازیابی کنند در رقابت با بازیابی اطلاعات توسط انسان از کارآیی بهتری برخوردار می باشند . بسیاری از روشهای هوش مصنوعی در بازیابی اطلاعات آزمایش شده و شبکه های عصبی به طور اخص ، به نظر می رسد که دارای خصوصیات لازم برای هوشمند تر نمودن فنون بازیابی اطلاعات باشند . پیاده سازی شبکه عصبی با ابعاد فضای بزرگتر که معمولاً در بازیابی اطلاعات به این فضای بزرگ اشاره می شود ، از لحاظ سخت افزاری گران بوده و پیاده سازی این گونه شبکه ها مستلزم صرف زمان زیاد می باشد .

۳-۱ استفاده از شبکه های عصبی در بازیابی اطلاعات

در مدلهای مختلف شبکه های عصبی ، اطلاعات به صورت شبکهٔ وزن دار نمایش داده می شود . بر خلاف تکنیکهای سنتی پردازش اطلاعات ، مدلهای شبکه های عصبی به عنوان خود- پردازشگر[1] بدون دخالت برنامه خارجی دیگر در شبکه ، عمل می نماید .

شبکه با رفتار هوشمند خود در فعل و انفعال های محلی که به طور همزمان میان اجزاء شبکه رخ می دهد ، به پردازش داده ها می پردازد . مطابق نظریه Lin مدلهای شبکه عصبی با مدلهای پردازش اطلاعات سنتی حداقل در دو روش زیر با یکدیگر تفاوت اساسی دارند :

[1] – Self- processing

۱. خود- پردازش شبکه های عصبی : مدلهای پردازش اطلاعات سنتی به طور عمده از ساختار داده هایی که همیشه بوسیله یک سؤال خارجی بدست می آیند ، استفاده می کنند . در شبکه های عصبی ، گره ها و پیوندها به عنوان یک عامل[1] پردازشگر فعال عمل می نمایند و بطور کلی هیچ عامل فعال خارجی که در رفتار گره ها و پیوندهای شبکه تأثیر گذارد، وجود ندارد .

۲. مدلهای شبکه عصبی، که در ارتباط با عناصرپیچیده رفتار کلّ سیستم را که از ترا کنشهای همزمان محلی بر شبکه ناشی می شود ، نمایش می دهد .

در مدل بازیابی اطلاعات سنتی ، پردازش خارجی که بر روی ساختار داده ها عمل می کند معمولاً به تمامی مجموعه قوانین شبکه دسترسی کلی دارد و پردازش عمدتاً ترتیبی می باشد .

محاسبات شبکه های عصبی به نظر می رسد که در مقایسه با مدل فضای برداری و مدلهای احتمال تناسب بهتری با مدلهای بازیابی سنتی داشته باشد .

از مزایای استفاده از شبکه عصبی در بازیابی اطلاعات می توان به موارد زیر اشاره نمود :

۱. در زمانی که اطلاعات (کلید واژه) مورد جستجو دقیقاً در مدارک پیدا نشود با استفاده از شبکه عصبی می توان به بازیابی داده هایی که از نظر همسایگی نزدیکتر به اطلاعات خواسته شده هستند ، پرداخت .
۲. دسته بندی اطلاعات با الگوهای مشترک
۳. تغییر اطلاعات ذخیره شده در پاسخ به درخواست جدید کاربر

۳. پیشینه تحقیق

داس کاکس و دیگران (۱۹۹۰) مرور جامعی درباره کاربرد مدلهای ارتباطی در بازیابی اطلاعات انجام داده اند . بخش مهمی از تحقیقات پیرامون بازیابی اطلاعات را می توان در چارچوب مدلهای ارتباطی مورد توجه قرار داد . برای مثال ، از آنجا که تمام مدلهای ارتباطی به عنوان

[1] -Agent

سیستمهای رده بندی ورودی – به- خروجی مطرح اند ، خوشه بندی مدرک را می توان به عنوان رده بندی فضای مدرک * مدرک در نظر گرفت. ساخت اصطلاحنامه به عنوان سیستمی هماهنگ با فضای نمایه * نمایه مطرح بوده و جستجو را نیز می توان به عنوان و ارتباط پیوند در فضای مدرک * نمایه تلقی نمود .

وانگ، کای ، و یائو (۱۹۹۳) روشی را برای محاسبه کلمات مرتبط با استفاده از شبکه feed forward سه لایه ای با تابع threshold خطی پیشنهاد کردند. گره ها در لایه پنهان، نمایانگر کلمات "پرس و جو" بوده و لایه خروجی فقط شامل یک گره است که ورودی را از همه کلمات "پرس و جو" می گیرد . کلمه مربوط توسط پیوندهای وزن دار که نرونهای مختلف را با یکدیگر متصل می نماید مدلسازی شد و شبکه توسط الگوریتم یادگیری پرسپترون بدون نیاز به معرفی پارامترهای ویژه آموزش داده شد .

کراستانی و ریسبرگن (۱۹۹۳) در مقاله خود مدلی از شبکه را ارائه می دهند که می تواند در ایجاد طرح مفهومی و منطقی جهت کاربردهای بازیابی اطلاعات مورد استفاده قرار گیرد . این مدل دارای خصیصه های انعطاف پذیر قابل توجهی است که می توان آن را به طرق مختلف و مؤثر بکار گرفت . این نویسندگان در سال ۱۹۹۷ شبکه تبدیل [1] را برای بهینه سازی پرس و جو پیشنهاد کردند . این شبکه شامل شبکه پس انتشار خطا با یک یا چند لایه پنهانی است که در آن ورودی و خروجی طرح واره هایی از بازنمونها بحساب می آید.

گرونفلد (۱۹۹٦) با استفاده از شبکه عصبی هاپفیلد ، گره هایی را برای مفاهیم "پرس و جو" و نیز گره هایی را برای " مدارک " در نظر گرفت . مدارکی که به این ترتیب بیشتر فعال می گردید برای بازیابی انتخاب می شدند . گرونفلد با استفاده از این مدل پیوند بین گره ها را بر اساس ماتریس مدرک * کلمه که به وسیله الگوریتم مشترک نمایه تعریف می شود برقرار ساخت.

[1] - Transformation Network

هاتانو و دیگران (۱۹۹۷) در ارتباط با خوشه بندی مؤثر و بازیابی متن و داده های ویدیوئی مبتنی بر شباهتهای موجود ، نظام سامان دهنده اطلاعات را پیشنهاد کردند . به جای کلید واژه ها این نویسندگان از مدل فضای برداری و کد گذاری تصویر DCT [1] به منظور استخراج خصایص داده ها استفاده نمودند . داده ها بر حسب شبکه عصبی کوهنن (SOM) خوشه بندی گردیده و نتیجه به یک شکل سه بعدی نمایش داده می شود.

هاتانو و همکاران وی معتقدند که سیستم پیشنهادی آنان به طور مؤثری به استفاده مجدد از امتیازات داده های تصویر و متنی توزیع یافته کمک می کند.

مقایسه میان الگوریتمهای SOFM که بر اساس بسامد واژه ها استوارند و الگوریتمهایی که بر اساس اندازه گیری Salton هستند ، نشان می دهد که الگوریتمهای SOFM جهت خوشه بندی اسناد و تولید نگاشت کلی اسناد مؤثرتر باشند . ماندل (۱۹۹۸) از شبکه عصبی "پس انتشار خطا" [۲] برای ساخت مدل کاسی میر [۳] به منظور تطبیق میان بازنمودن پرس و جو و مدرک استفاده نمود . پرس و جو و مدرک هر دو ، به عنوان ورودی شبکه عمل می کنند و شبکه به مشابهت آنها در لایه خروجی می پردازد . به عبارت دیگر ، هر چه ضریب شباهت بیشتر باشد میزان ربط بین مدرک و پرس و جو بیشتر خواهد بود .

SOFM همچنین در پروژه [۴] DLI در دانشگاه ایلینویز نیز به کار گرفته شد .

چانگ و همکاران او (۱۹۹۸) در مقاله ای به عنوان " نمایه سازی موضوعی خودکار با استفاده از شبکه عصبی اشتراکی" سیستم "تخصیص گر مفهوم " [۵] را که یک سیستم نمایه سازی موضوعی خودکار مبتنی بر شبکه هاپفیلد است پیشنهاد نمودند . در این سیستم ، از مجموعه ای از اسناد استفاده می شود که به طور خودکار زیر مجموعه ای از واژه های

[1] -Discrete Cosine Transform
[2] - Back propagation
[3] -Cognitive SIMilarity Learning in IR (COSIMIR)
[4] - Digital Library Initiative
[5] -Concept Assigner

موضوعی را که "فضای مفهوم" [1] نامیده می شود ، ایجاد می نماید . برای نمایه سازی خودکار یک مدرک خاصّ ، مفاهیم استخراج شده از مدرک مزبور، در فضای برداری به ورودیهای شبکه هاپفیلد تبدیل می شوند . مدلی از سیستم نمایه سازی موضوعی خودکار این نویسندگان به عنوان بخشی از پروژه Interspace که یک محیط نمایه سازی و بازیابی است به اجرا در آمده است . این سیستم از نمایه سازی معنایی آماری پشتیبانی می کند .

۳. روش پژوهش

داده های کتابخانه منطقه ای از نوع متن بوده و اطلاعات درخواستی از طریق کلید واژه ها قابل دسترسی می باشند . جهت استفاده از این داده ها ، برای نمونه از داده های موجود در دیسک نوری INIS
(International Nuclear Information System)
استفاده شد. موضوع مورد بحث cross-section در سه مقوله زیر می باشد :

1. elastic scattering cross-section
2. absorbtion
3. fission

از هر مقوله۵۰ مدرک و در مجموع ۱۵۰ مدرک بررسی شد . در سه مقوله فوق، چهار کلید واژه Neutron , Proton , Electron , Positron در نظر گرفته شد.

هدف، استفاده از الگوریتم خوشه بندی شبکه عصبی جهت دسته بندی سه رده elastic , absorbation و fission است . با اطلاع از اینکه در مدارک مورد بحث در زمینه fission کلید واژه Neutron بیش از سه کلید واژه دیگر بوده و در زمینه elastic کلید واژه Proton بیشتر و در زمینه absorbation کلید واژه های Electron و Neutron بیشتر است ، به محاسبه وزن مدارک پرداخته شده و کدهایی جهت پیش پردازش و رمز گذاری مدارک متنی به بردارهای عددی نوشته شد . جهت تعیین بردار وزنی هر مدرک ، از بسامد کلید واژه در مدرک استفاده شد به عبارت دیگر وزن w_{ik} مدرک به صورت بسامد

[1] -Concept Space

کلید واژه یا کلمه t_k در مدرک d_i تعریف می شود . از فرمول زیر جهت تعیین وزن استفاده شد:

$$W_{ik} = (tf_{ik} . \log(N/n_k) / \sqrt{ \sum_{j=1}^{t} (tf_{ij})^2 . (\log (N/n_j))^2 } \qquad (۲)$$

در این فرمول، tf_{ik} بسامد کلید واژه t_k در مدرک d_i می باشد، و پارامتر N تعداد مدارک را نشان می دهد، و n_k بیانگر تعداد مدارکی است که شامل واژه یا کلید واژه t_k می باشند. برای ۱۵۰ مدرک نمونه پارامترهای فوق معین شد و نتیجه در پیوست این گزارش ارائه شده است.

الگوریتم شبکه عصبی جهت خوشه بندی مدارک با اختصاص یک گره برای هر خوشه در شبکه پیاده سازی می شود . هر گره، اندازه شباهت میان مدرک موجود و مرکز ثقلی که خوشه به همراه گره است، را نشان می دهد .

ابتدا ، ضریب شباهت میان مدرک ورودی و مرکز ثقل خوشهٔ موجود محاسبه می شود . اگر ضریب شباهت S_1 بزرگتر از آستانه S_{1avg} باشد در آن صورت گره ورودی فعال می شود و سپس یک حلقه بازگشتی جهت اختصاص مدرک ورودی به خوشه ایجاد می شود . گره هایی که به مدرک نزدیک نباشد غیر فعال می شوند . در مرحله دوم، تمام گره هایی که به عنوان گره برنده انتخاب شده بودند جهت محاسبهٔ ضریب شباهت انتخاب می شوند تفاوت ضریب شباهت S_2 جهت اطمینان یافتن از اینکه خوشهٔ برنده شده به مدرک ورودی نزدیک است ، محاسبه می شود . اگر شباهت گره به خوشه زیاد باشد آنگاه گره به خوشه اضافه شده و مرکز ثقل روز آمد می شود در غیر اینصورت، خوشهٔ جدیدی برای مدرک ورودی جدید ساخته می شود . با استفاده از این داده ها شبکه مصنوعی هوشمندی که عمل دسته بندی داده ها را مطابق با یادگیری غیر تعلیم یافته انجام می دهد پیاده سازی می شود. فرایند یادگیری SOM را می توان به عنوان یادگیری رقابتی در نظر گرفت . ایده اصلی یادگیری رقابتی، تنظیم یک خوشه آزمایشی(C) شبکه با بالاترین

سطح فعالیت مطابق با ورودیهای تصادفی انتخاب شده است . سطح فعال خروجی بر اساس فاصله اقلیدسی میان بردار وزن خوشه ها m_c و ورودی تعیین می شود . مدل فضای برداری برای نمـایش داده ها به وسیله بـــــــردارهایی با وزن W تعیین گردیده، به طوری که $\sum (W_i)^2 = 1$ باشد . شباهت میان دو واژه به وسیله محاسبه اندازه شباهت کسینوسی بردادها تعیین شد :

$$\frac{(w,v)}{\| W \|^2 \ \|V\|^2} = \frac{(w.v)}{| W | \ |V|} = \frac{(w.v)}{(1)\ (1)} = \sum W_i V_i \qquad (۳)$$

در این اندازه گیری، زاویه کسینوسی بین دو بردار معین گردید . یک بردار برای هر مدرک بر اساس کلید واژه های چهار گانه ساخته شد. الگوریتم پروژه در شکل ٥ نمایش داده شده است .

به طور کلی، چهار مرحله زیر جهت فرآیند یادگیری در نظر گرفته شد:

١. انتخاب تصادفی ورودی $x(t)$

٢. محاسبه فاصله میان بردارهای وزن و بردار ورودی با استفاده از فرمول زیر :

$$D_i(t) = \| x(t) - m_i(t) \| = \sqrt{\sum_{p=1}^{n} (\epsilon_p - \mu_{ip})^2} \qquad (٤)$$

m_i به بردار وزن i اشاره می نماید و $\| . \|$ نرم بردار اقلیدسی است .

٣ . خوشه برنده از فرمول زیر محاسبه می شود :

$$C : \| x(t) - m_i(t) \| = \min_i (D_i(t)) \qquad (٥)$$

٤ . تنظیم بردارهای وزن در همسایگی خوشه برنده .

لازم به یادآوری است که قانون یادگیری استفاده شده در این مبحث به شرح زیر می باشد :

$m_i(t+1) = m_i(t) + \mu(t) . h_{ci}(t) . [x(t) - m_i(t)]$ (٦)

در فرمول فوق $\mu(t)$ ضریبی است که با زمان کاهش می یابد و $h_{ci}(t)$ تابع همسایگی است که در شعاع همسایگی برنده متقارن می باشد .

```
For  N training epochs
For each trainig documents
        Word_index=0
        Neighbor_ Word_index= Word_index+1
        Adjust context vector for word(Word_index) and
word(neighbor_Word_index)
        D=w1-w2  where
                w1=vector for word(Word_index)
               w2= vector for word(neighbor_Word_index)
        w1 (k+1)=w1 (k)- μw k1 d   where
                μw = learning rate for word neighbor adjustments
                        k1 =(w1_update*(( neighbor_Word_index) –
                        (Word_index)))  -1
        w2 (k+1)=w2 (k)+ μw k2 d   where
                        k2 =(w2_update*(( neighbor_Word_index) –
                        (Word_index)))  -1
        normalize w1 and w2
        if (( neighbor_Word_index) – (Word_index)) <
        max_neighbor_word
        increment  neighbor_Word_index
else if not done with all words in document
         increment  Word_index
caculate vector for document
d=w-v where
        w= vector for the word
        v=vector for the entire document
 w(k+1)=w(k)- μd d
        where  μd  is learning rate for word-to-document
adjustment( μd << μw )
renormalize w
```

شکل ٥- الگوریتم پروژه

در این پروژه از تابع همسایگی گاوس ساده با فرمول زیر استفاده شد :

$$h_{ci}(t) = \exp\left(- \frac{\| r_c - r_i \|}{2 .\delta (t)^2} \right) \qquad (۷)$$

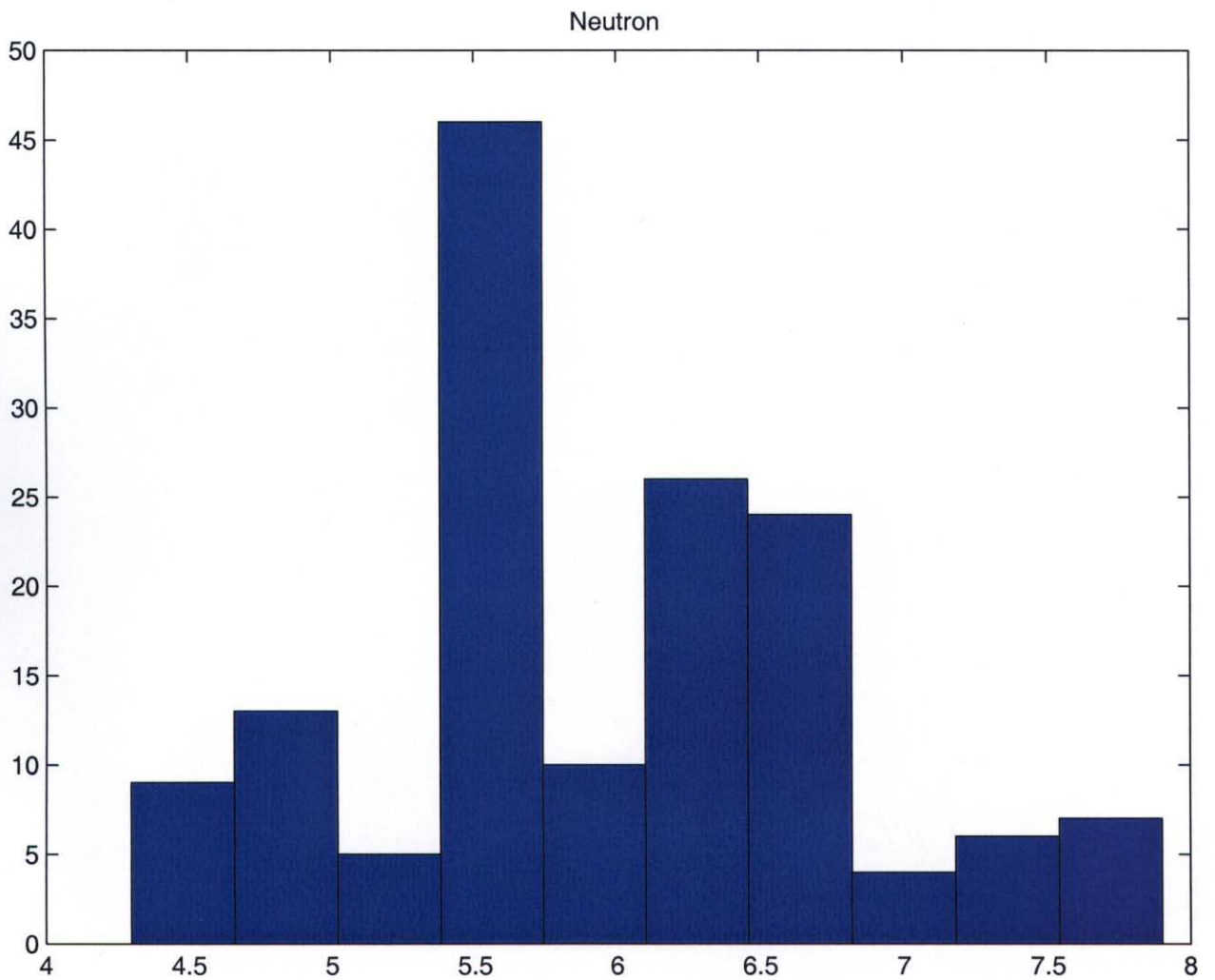
مقدار δ با افزایش تکرار زمان یادگیری t ، کاهش می یابد . هدف از به کار گیری این پارامتر بیان فاصله تابع همسایگی بر اساس واحدهای مؤثر می باشد . r به عنوان شعاع همسایگی در نظر گرفته شده است .

با کاهش میزان یادگیری و کاهش محدوده همسایگی ، فرآیند یادگیری به سوی حالت پایدار خود پیش می رود . حالت پایدار زمانی برقرار می شود که تغییرات زیادی در بردار های وزن مختلف مشاهده نشود . گرچه ، عمل فرایند یادگیری ممکن است زمانی که هر دادهٔ ورودی مکرراً به یک خوشه نسبت داده شود ، پایان یابد .

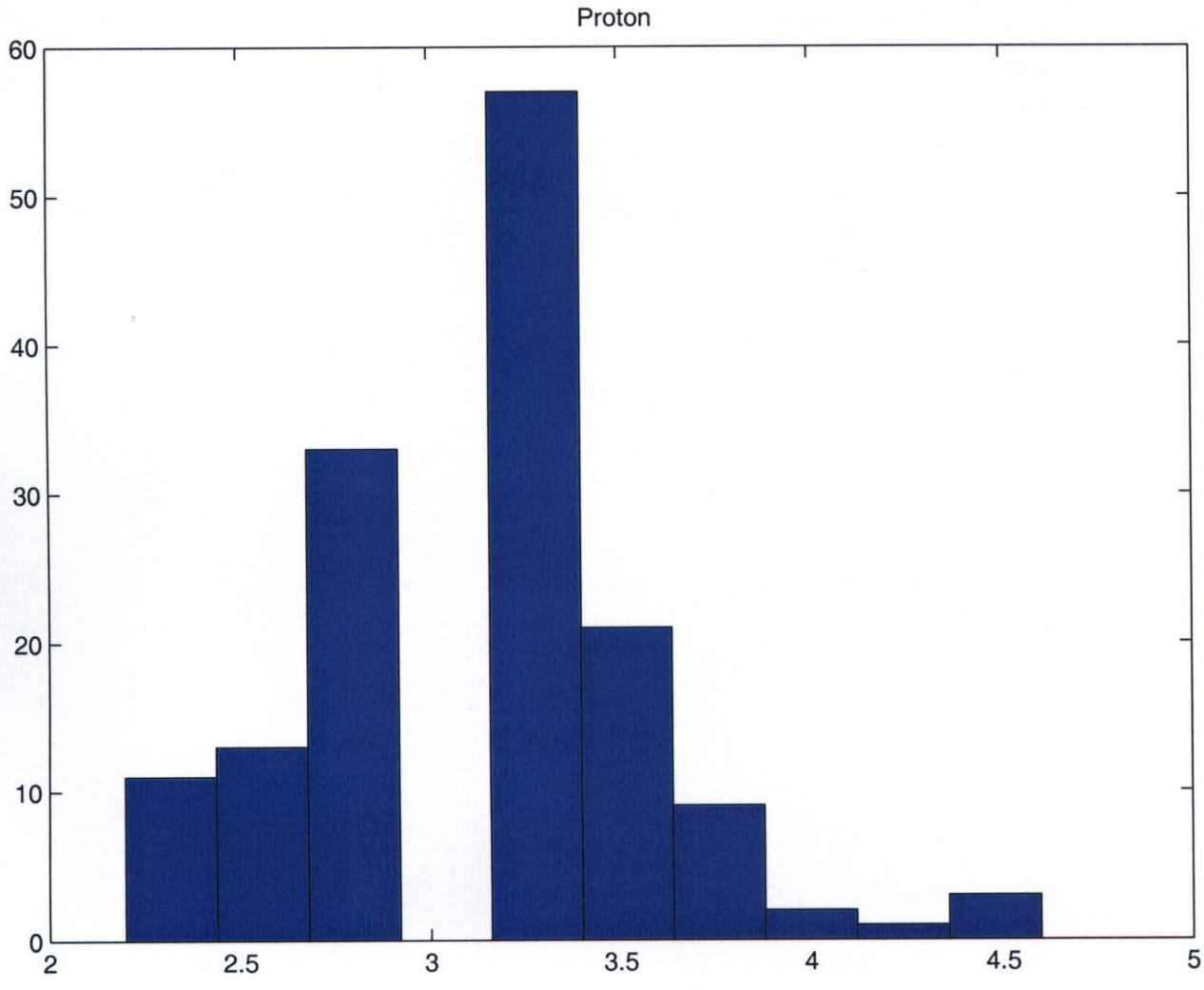
۱. نتایج

اجرای این پروژه با استفاده از نرم افزار MATLAB 6.2 در محیط Windows 2000 صورت گرفت . با استفاده از فرمول ۲، بردار وزن مدارک تعیین شد (پیوست ۱)، سپس هیستوگرامی از کلید واژه های موجود در مجموعه مدارک مطابق شکلهای ٦ تا ۹ تهیه شد.

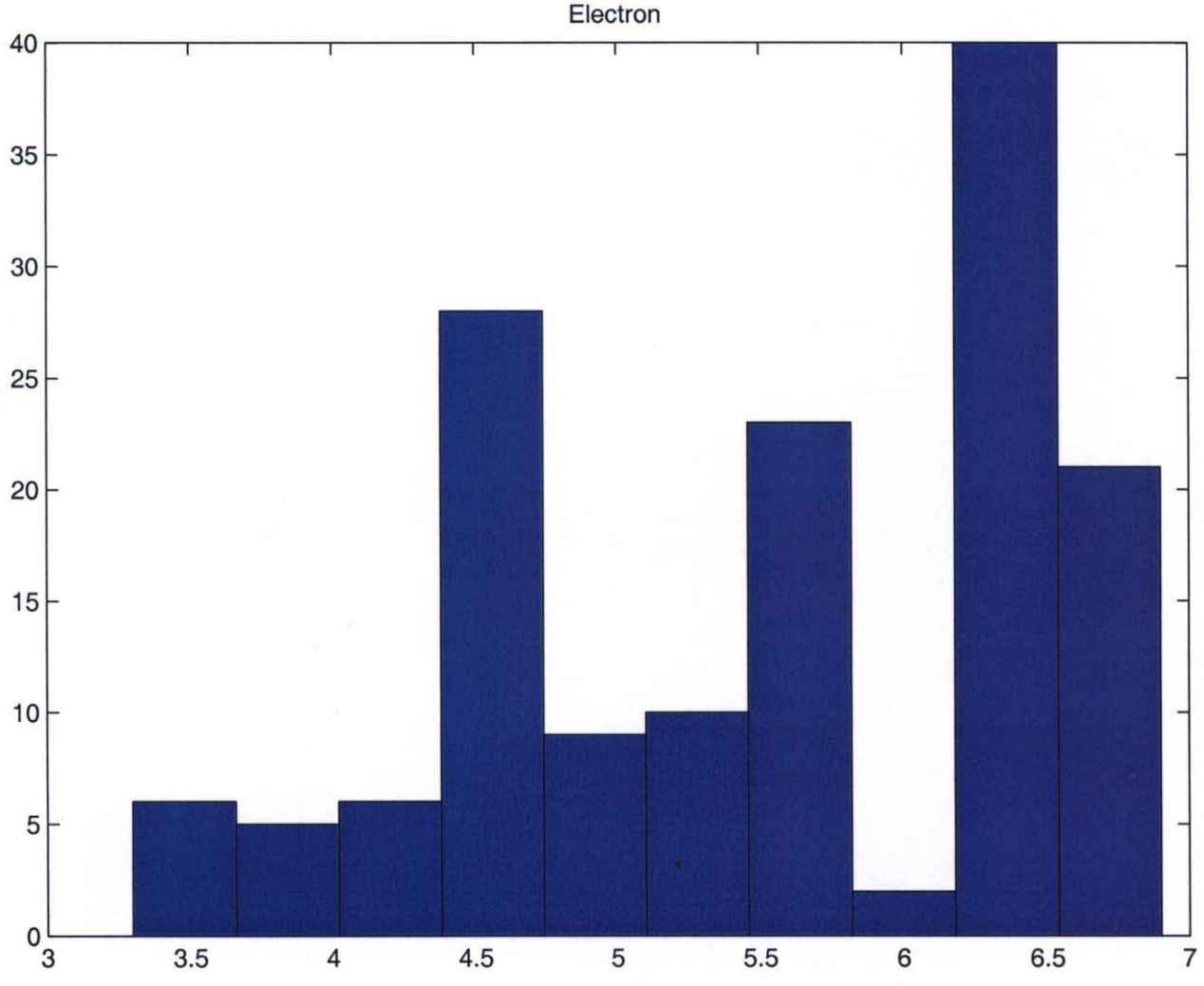
از آنجایی که الگوریتم SOM بر اساس فاصله اقلیدسی استوار است ، محدوده اندازه متغیرها در اختصاص متغیر به خوشه خاص ، بسیار مهم می باشد و معمولاً متغیرها نرمال شده، بطوری که هر جزء، دارای واریانس واحدی بوده و با داده های نرمال شده عمل یادگیری انجام می پذیرد .



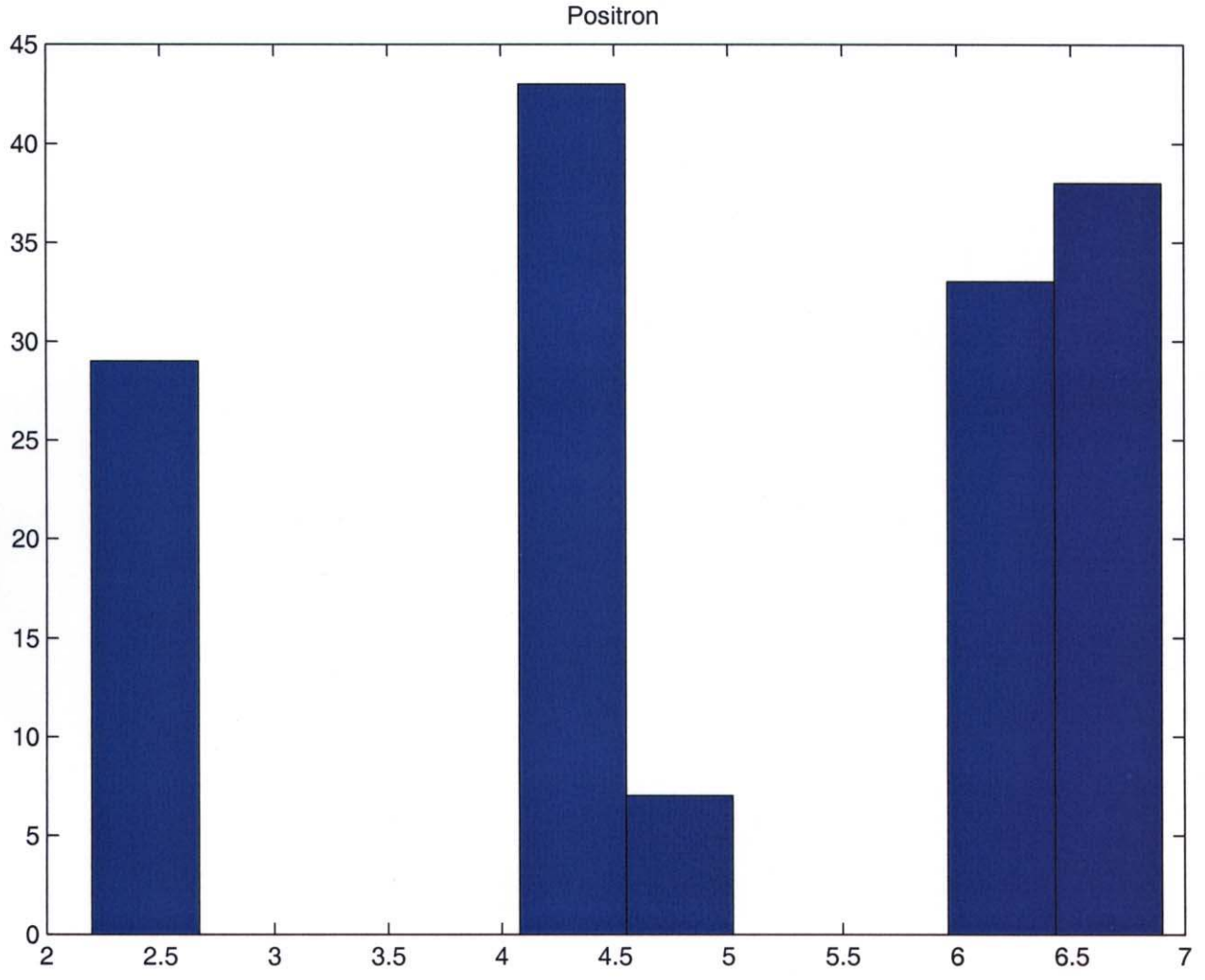
شکل ٦- هیستوگرام نوترون در مجموعه مدارک



شکل ۷- هیستوگرام پروتون در مجموعه مدارک



شکل ۸- هیستوگرام الکترون در مجموعه مدارک



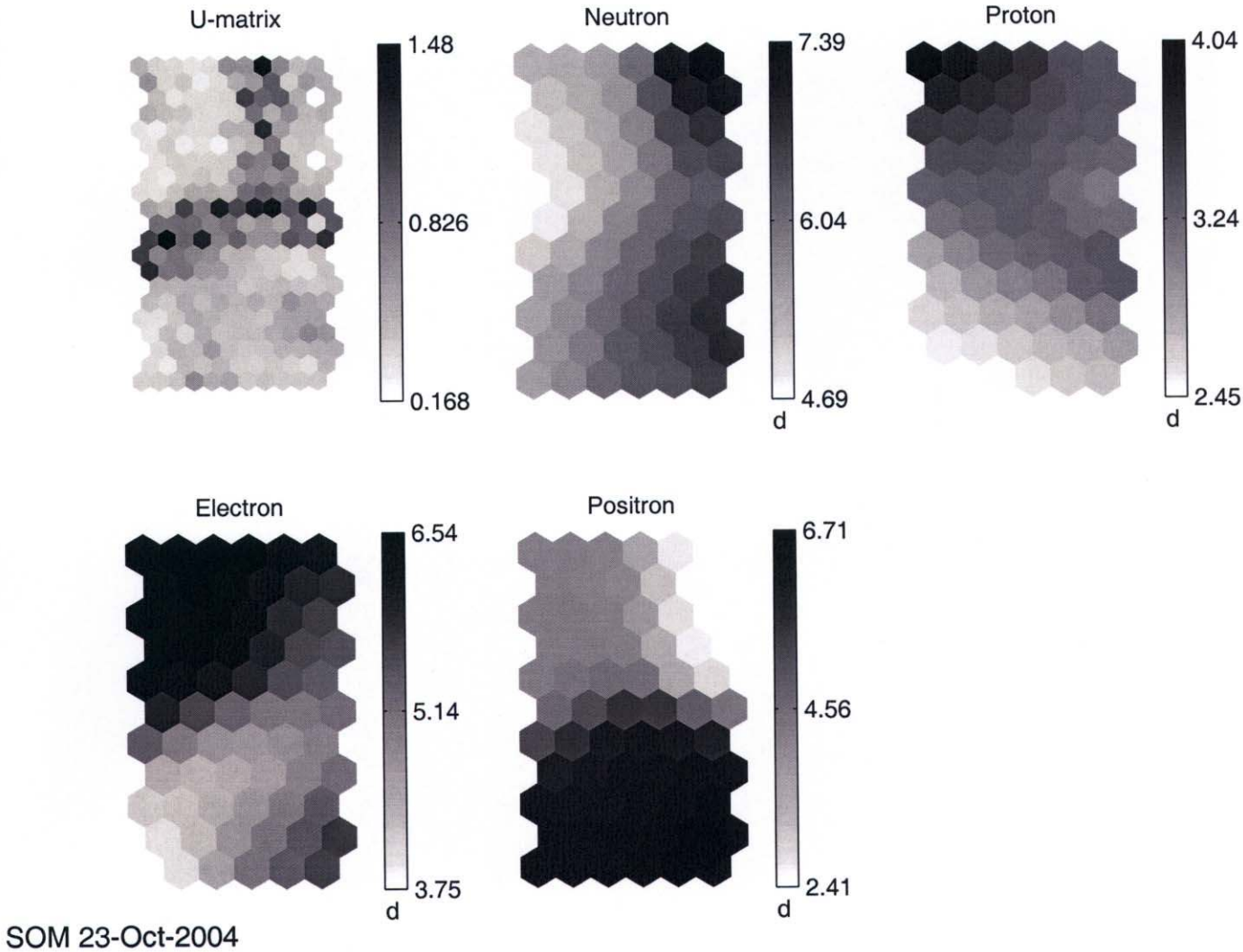
شکل ۹- هیستوگرام پوزیترون در مجموعه مدارک

مجموعه داده های مورد بررسی شامل ۵۰ نمونه از سه دسته مربوط به موضوع cross section انتخاب شد که نگاشت آن در شکل ۱۰ نشان داده شده است . همانطور که در این شکل مشاهده می شود ، محدوده مقدار متغیرهای نرمال شده هر کلید واژه در مجموعه مشخص شده است . ماتریس U در این شکل بیانگر فاصله بین همسایگی ها است و ساختار خوشه بندی شبکه SOM را مشخص می نماید. برای محاسبه ماتریسU از تمام یا تعدادی از متغیرهای شبکه استفاده می شود که در این طرح آزمایشی به علت کم بودن تعداد متغیرها، از تمام مقادیر برای محاسبه استفاده شد . مقادیر بیشتر در این ماتریس بیانگر فاصله همسایگی بین نگاشتها است و بنابراین محدوده خوشه را بیان می کند . خوشه ها، به طور نمونه، با مساحتهای یکنواختی از مقادیر کمتر نشان داده شده اند . با استفاده از نمودار ستونی رنگها مشاهده می شود که چه رنگی دارای چه مقداری است و اولین مقدار بیانگر مقدار متغیر در ساختار شبکه SOM است .

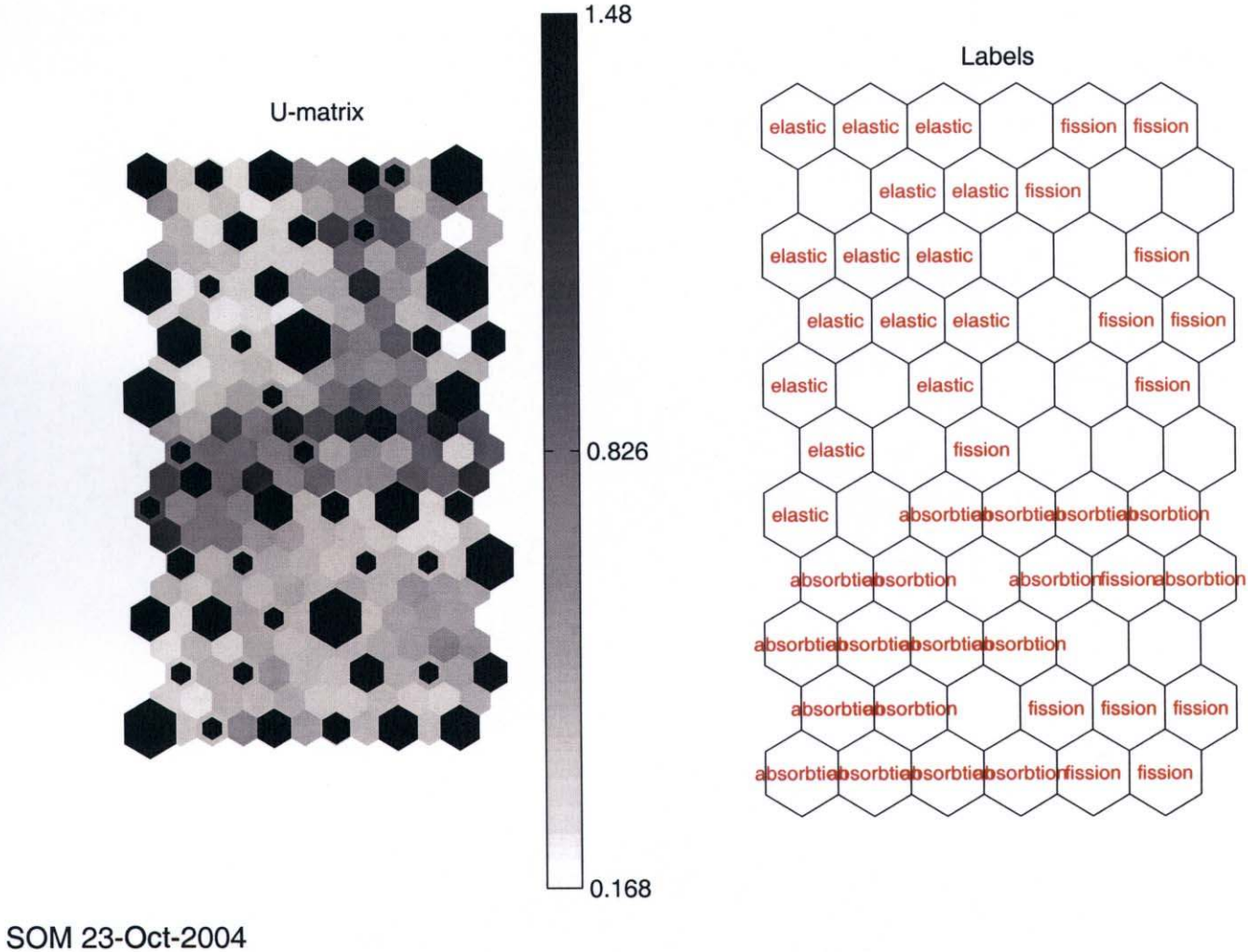
در شکل ۱۱ ماتریس U به همراه بر چسب آنها نمایش داده می شود . این شکل بیانگر دسته بندی خوشه ها است .

شکل ۱۲ ماتریس فاصله را به صورت نمایش سطحی نشان می دهد . مقادیر در محور Z نشانگر متوسط فاصله به نرونهای همسایگی نگاشت است . این ماتریس بسیار نزدیک به ماتریس U است . توسط این شکل روابط توپولوژی ماتریس به صورت نمایش سطحی به راحتی قابل تشخیص است .

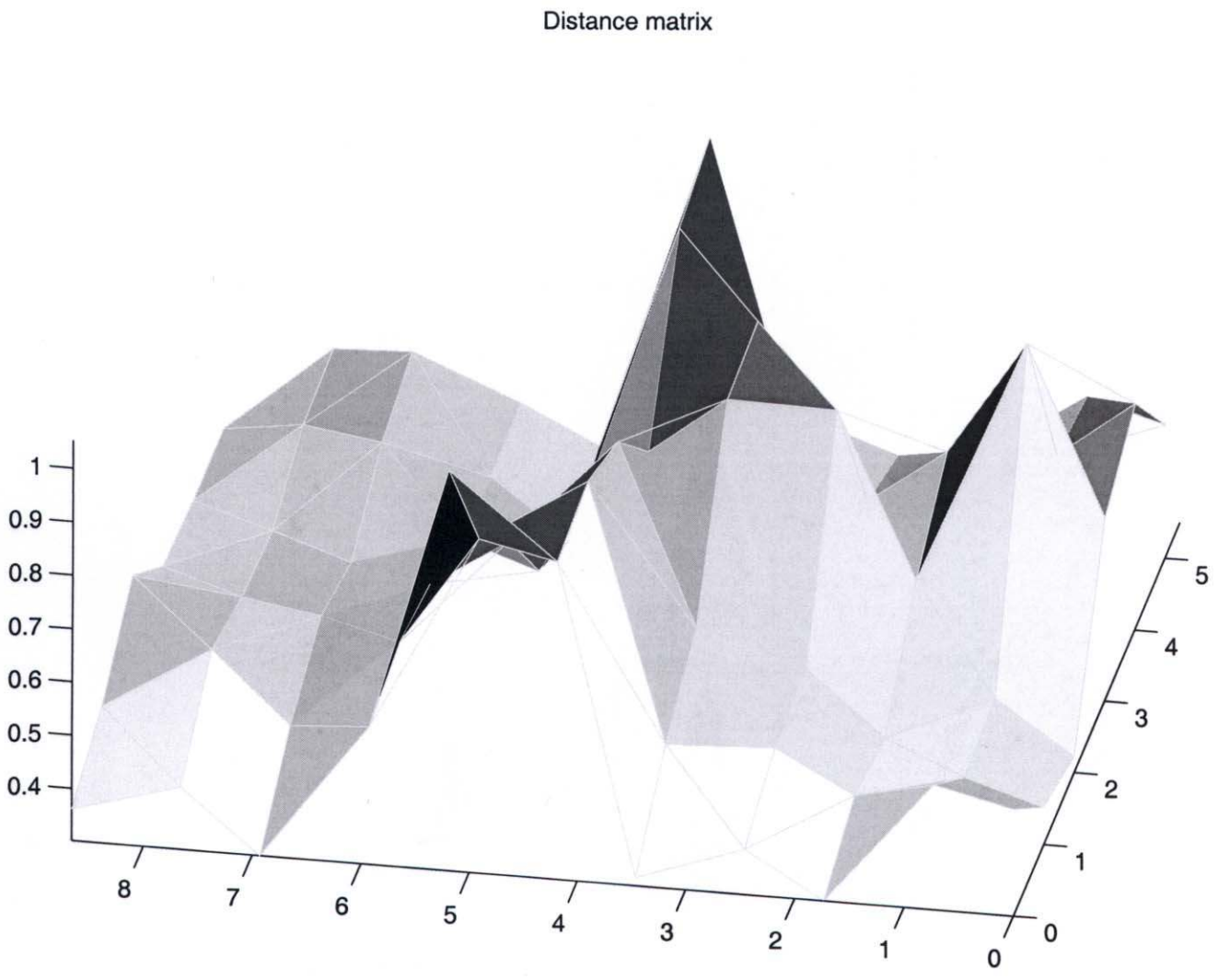
شکل ۱۳ الگویی از شبکه SOM طراحی شده را با نرونهای آن در فضای سه بعدی مشخص ساخته و شکل ۱٤ شبکه SOM را به همراه نمایش سه بعدی داده ها نشان می دهد .



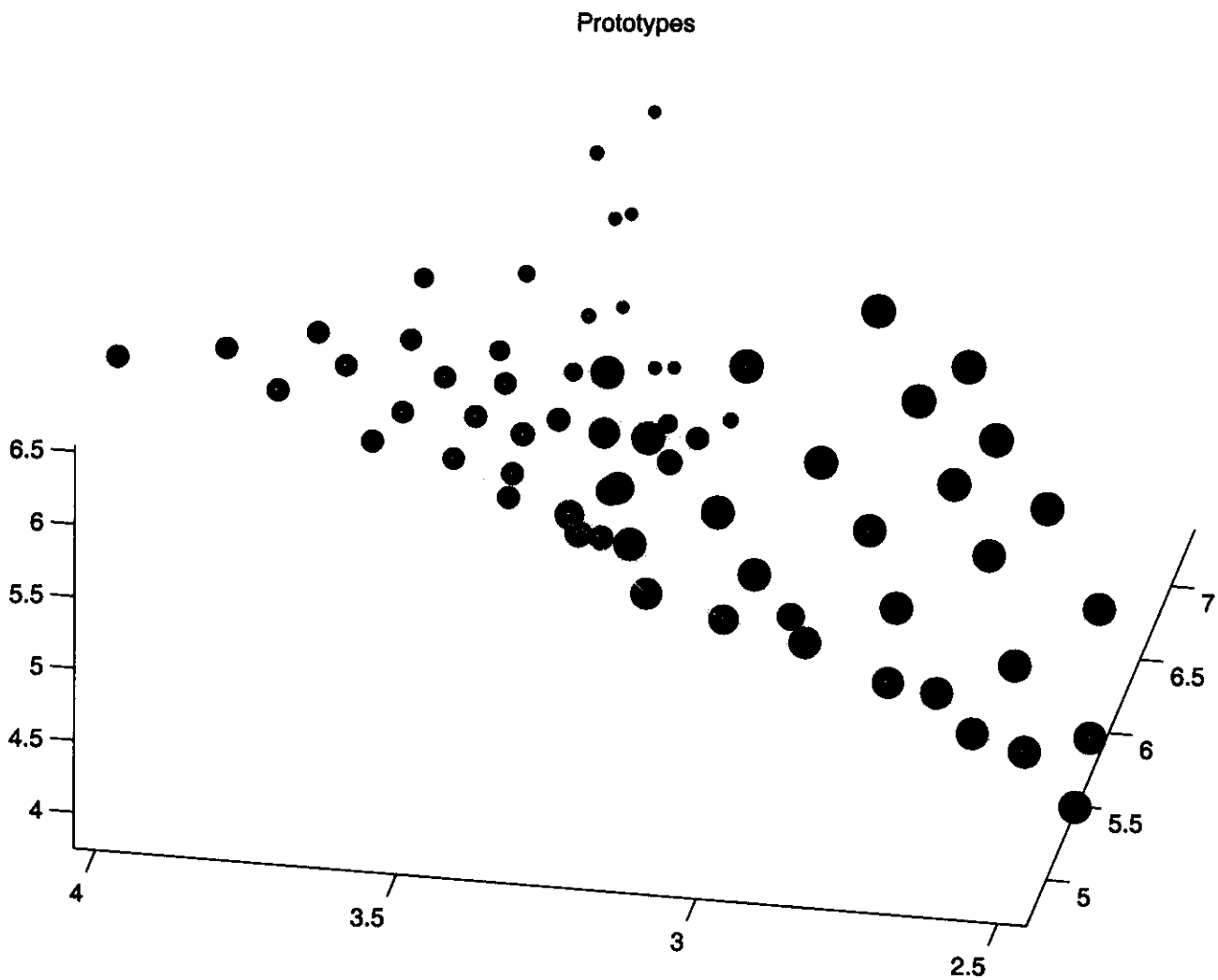
شکل ۱۰- مقدار متغیرها در هر نگاشت به همراه ماتریس U



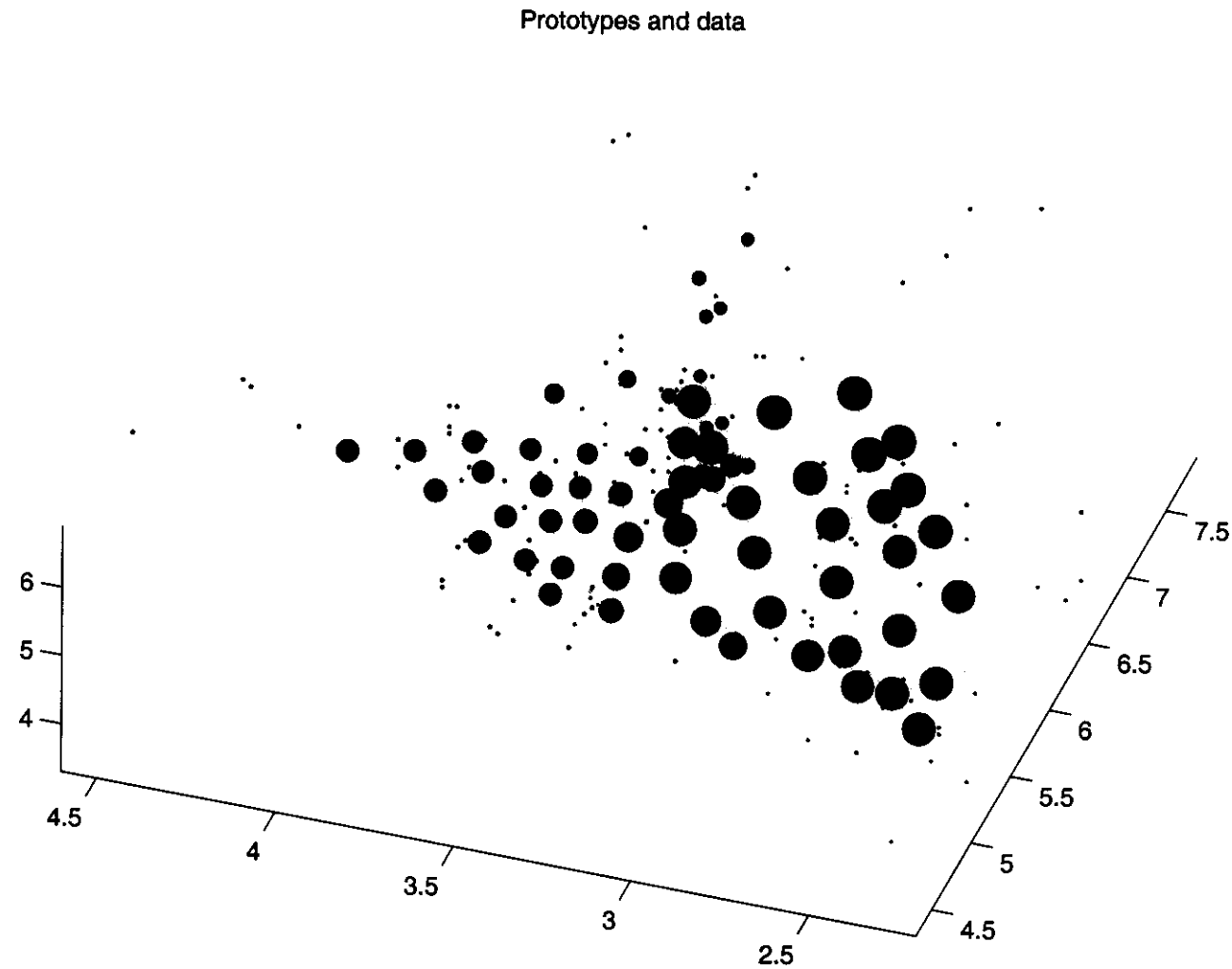
شکل ۱۱- ماتریس Uبه همراه نمایش خوشه ها



شکل ۱۲- نمایش صفحه ای ماتریس فاصله



شکل ۱۳- شبکه SOM



شکل ١٤- شبکه SOM به همراه داده ها

جهت ارزیابی سیستم ، مقادیر دقت بازیافت[1] و بازیافت[2] محاسبه گردید. بر اساس تعریف، مقدار دقت بازیافت برابر است با نسبت تعداد مدارک مرتبط بازیابی شده به تعداد کل مدارک بازیابی شده و مقدار بازیافت برابر است با تعداد مدارک مرتبط بازیابی شده به تعداد کل مدارک مرتبط در مجموعه . مقدار دقت بازیافت میان 0.2 و 0.3 است و مقدار بازیافت حدود 0.85 است . هر چه مقدار بازیافت به یک نزدیکتر باشد کارایی سیستم در بازیابی اطلاعات بالاتر است بنابراین مقدار بدست آمده عدد قابل قبولی است . شکل ۱۵ دقت بازیافت و شکل ۱٦ رابطه بین دقت بازیافت و بازیافت را نشان می دهد .

۵ . بحث و نتیجه گیری

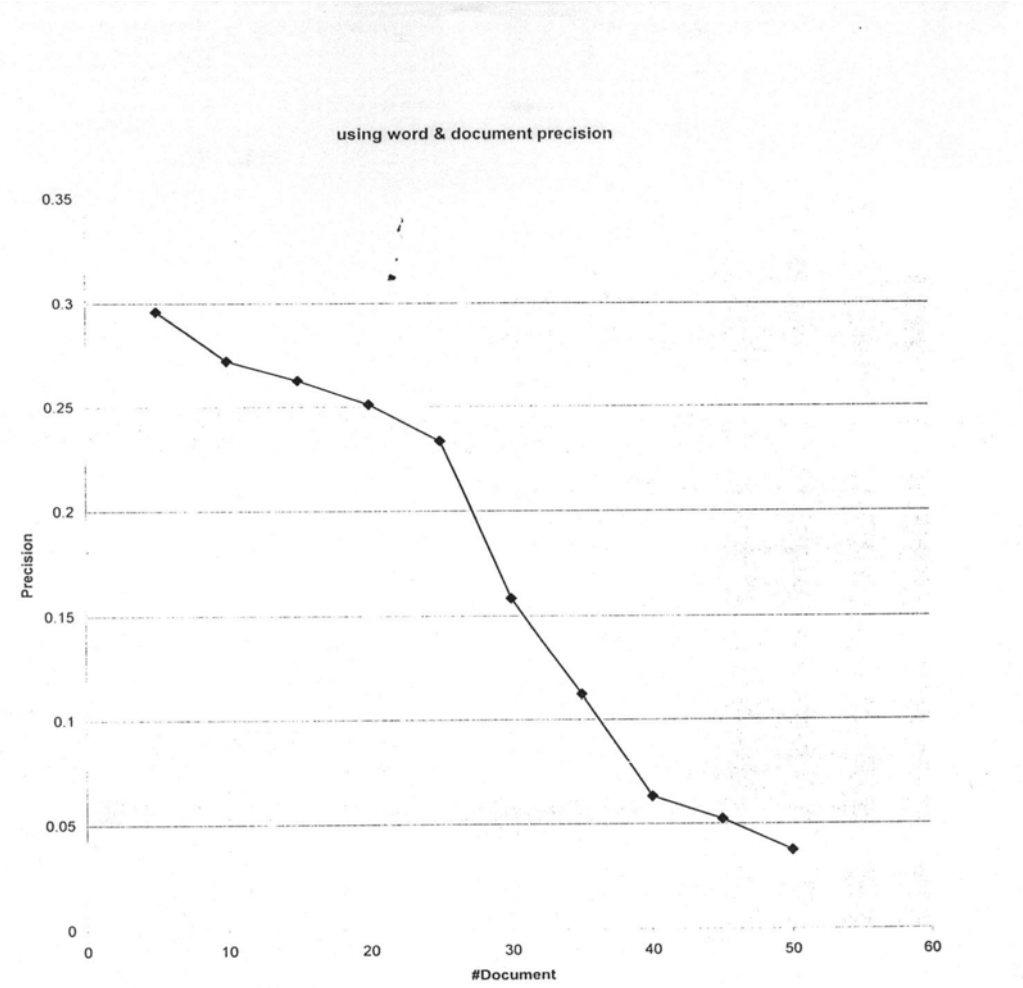
در این طرح با توجه به نوع داده های موجود در کتابخانه منطقه ای که از نوع متنی می باشد به شناخت و بررسی شبکه عصبی مناسب با نوع داده ها در بازیابی اطلاعات پرداخته شد . شبکه های عصبی هاپفیلد و پس انتشار خطا نیز مطالعه گردید که با توجه به خصوصیت داده های متنی و نیز نحوه بازیابی اطلاعات ویژگیهای شبکهSOM بسیار نزدیک با هدف طرح پژوهشی دیده شد و از این شبکه جهت این طرح استفاده شد . هدف این طرح پژوهشی شناسایی شبکه عصبی مؤثر جهت بازیابی اطلاعات متنی است لذا در مقیاس کوچک این شبکه طراحی گردید. در عمل جهت بازیابی کل اطلاعات موجود بایستی مراحل زیر انجام پذیرد :

۱. گروه بندی موضوعی مدارک

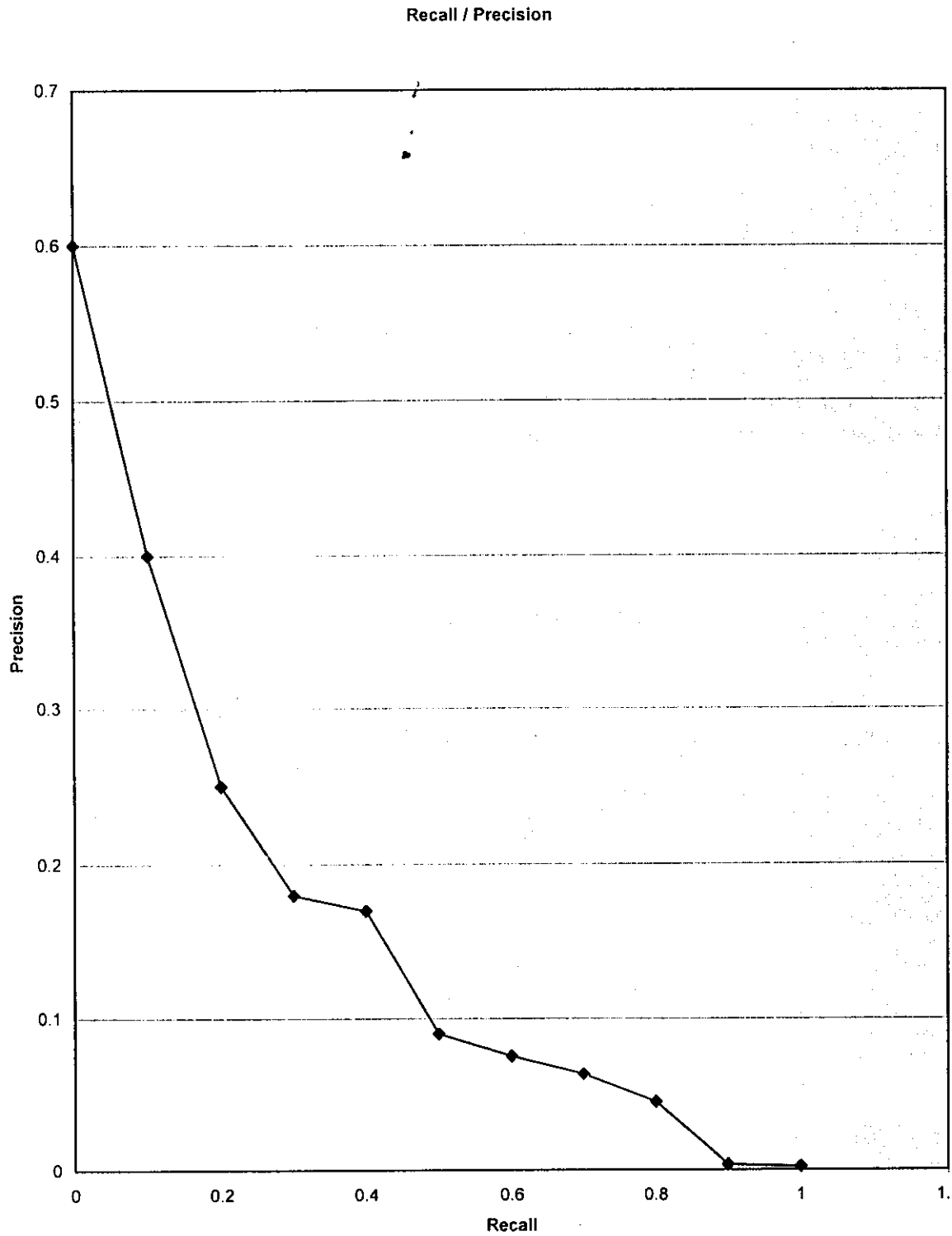
جهت دسته بندی مدارک معمولاً از سیستمهای نمایه سازی خودکار متن برای تخصیص گروه های موضوعی به مدارک متنی استفاده می شود . فایده این دسته بندی در آن است که با اختصاص یک مدرک در یک گروه

[1] -Precision
[2] -Recall

موضوعی خاص، جستجو محدود شده و سرعت بازیابی اطلاعات بیشتر می شود . جهت دسته بندی مدارک می توان از شبکه عصبی استفاده کرد.



شکل ۱۵– نمودار دقت بازیافت



شکل ١٦- رابطه بین دقت بازیافت و بازیافت

به دلیل اینکه فضای خصیصه های مدارک متنی از ابعاد زیادی برخوردار هستند ، آموزش شبکه عصبی با اطلاعات خام با ابعاد زیاد، بسیار طولانی و کند می باشد. جهت بهبود این مسئله پیشنهاد می شود که از تکنیکهای کاهش فضای خصیصه ای داده ها مانند تکنیکهای CF-DF و DF و IF × IDF به منظور کاهش ابعاد جهت دسته بندی توسط شبکه عصبی استفاده کرد .

۲. شناسایی ویژگیهای مدارک متنی شامل مراحل زیر است :
   - استخراج کلمات
   - حذف واژه های غیر مجاز[1]
   - در نظر گرفتن ریشه کلمات
   - محاسبه بردار وزنی کلمات
   - تهیه فرهنگ لغات مناسب

۳. محاسبه بردار ویژگی ورودی برای هر مدرک متنی

٤. به کارگیری از الگوریتم ارائه شده در شکل ٥ جهت یادگیری شبکه عصبی SOM

با در نظر گرفتن خصوصیات شبکه های عصبی ، به نظر می رسد که این تکنیکِ هوش مصنوعی جهت بازیابی اطلاعات مؤثر باشد . در حال حاضر، در مدلهای کاربردی شبکه های عصبی در بازیابی اطلاعات تحقیقاتی صورت گرفته است و در آینده با پیشرفت بیشتر سخت افزار و نرم افزار ، به نظر می رسد که به سرعت بتوان از شبکه های عصبی استفاده های مؤثرتری نمود . در آینده با ارزانتر شدن سخت افزار مورد نیاز مدلهای شبکه عصبی امکان استفاده از شبکه عصبی ، با استفاده موازی جهت طبقه بندی مدارک امکان پذیر گشته و بازیابی اطلاعات با استفاده از شبکه عصبی بسیار سریع تر خواهد بود. همچنین، با پیشرفتهای نرم افزاری در ایجاد روشهای جدید کنترل توابع شبکه ، به منظور پیاده سازی مدلهای شبکه عصبی، تحولی مهم در سرعت بازیابی اطلاعات فراهم می شود .

[1] - Stop Word

پیوست

Neutron Proton Electron Positron
5.6 3.5 6.4 4.2 elastic
4.9 3.4 6.4 4.2 elastic
4.7 3.2 6.3 4.2 elastic
4.6 3.6 6.5 4.2 elastic
5.4 3.6 6.4 4.2 elastic
5.4 3.9 6.7 4.4 elastic
4.6 3.4 6.4 4.3 elastic
5.4 3.4 6.5 4.2 elastic
4.4 2.9 6.4 4.2 elastic
4.9 3.6 6.5 4.6 elastic
5.4 3.7 6.5 4.2 elastic
4.8 3.4 6.6 4.2 elastic
4.8 3.4 6.4 4.6 elastic
4.3 3.4 6.6 4.6 elastic
5.8 4.4 6.2 4.2 elastic
5.7 4.4 6.5 4.4 elastic
5.4 3.9 6.3 4.4 elastic
5.6 3.5 6.4 4.3 elastic
5.7 3.8 6.7 4.3 elastic
5.6 3.8 6.5 4.3 elastic
5.4 3.4 6.7 4.2 elastic
5.6 3.7 6.5 4.4 elastic
4.6 3.6 6.4 4.2 elastic
5.6 3.3 6.7 4.5 elastic
4.8 3.4 6.9 4.2 elastic
5.4 3.4 6.6 4.2 elastic
5.4 3.4 6.6 4.4 elastic
5.2 3.5 6.5 4.2 elastic
5.2 3.4 6.4 4.2 elastic
4.7 3.2 6.6 4.2 elastic
4.8 3.6 6.6 4.2 elastic
5.4 3.4 6.5 4.4 elastic
5.2 4.6 6.5 4.6 elastic
5.5 4.2 6.4 4.2 elastic
4.9 3.6 6.5 4.6 elastic
5.4 3.2 6.2 4.2 elastic
5.5 3.5 6.3 4.2 elastic
4.9 3.6 6.5 4.6 elastic
4.4 3.4 6.3 4.2 elastic
5.6 3.4 6.5 4.2 elastic
5.4 3.5 6.3 4.3 elastic
4.5 2.3 6.3 4.3 elastic
4.4 3.2 6.3 4.2 elastic
5.4 3.5 6.6 4.6 elastic
5.6 3.8 6.9 4.4 elastic

```
4.8 3.4 6.4 4.3 elastic
5.6 3.8 6.6 4.2 elastic
4.6 3.2 6.4 4.2 elastic
5.3 3.7 6.5 4.2 elastic
5.4 3.3 6.4 4.2 elastic
7.4 3.2 4.7 6.4 absorbtion
6.4 3.2 4.5 6.5 absorbtion
6.9 3.6 4.9 6.5 absorbtion
5.5 2.3 4.4 6.3 absorbtion
6.5 2.8 4.6 6.5 absorbtion
5.7 2.8 4.5 6.3 absorbtion
6.3 3.3 4.7 6.6 absorbtion
4.9 2.4 3.3 6.4 absorbtion
6.6 2.9 4.6 6.3 absorbtion
5.2 2.7 3.9 6.4 absorbtion
5.4 2.4 3.5 6.4 absorbtion
5.9 3.4 4.2 6.5 absorbtion
6.4 2.2 4.4 6.4 absorbtion
6.6 2.9 4.7 6.4 absorbtion
5.6 2.9 3.6 6.3 absorbtion
6.7 3.6 4.4 6.4 absorbtion
5.6 3.4 4.5 6.5 absorbtion
5.8 2.7 4.6 6.4 absorbtion
6.2 2.2 4.5 6.5 absorbtion
5.6 2.5 3.9 6.6 absorbtion
5.9 3.2 4.8 6.8 absorbtion
6.6 2.8 4.4 6.3 absorbtion
6.3 2.5 4.9 6.5 absorbtion
6.6 2.8 4.7 6.2 absorbtion
6.4 2.9 4.3 6.3 absorbtion
6.6 3.4 4.4 6.4 absorbtion
6.8 2.8 4.8 6.4 absorbtion
6.7 3.4 5.4 6.7 absorbtion
6.4 2.9 4.5 6.5 absorbtion
5.7 2.6 3.5 6.4 absorbtion
5.5 2.4 3.8 6.6 absorbtion
5.5 2.4 3.7 6.4 absorbtion
5.8 2.7 3.9 6.2 absorbtion
6.4 2.7 5.6 6.6 absorbtion
5.4 3.4 4.5 6.5 absorbtion
6.4 3.4 4.5 6.6 absorbtion
6.7 3.6 4.7 6.5 absorbtion
6.3 2.3 4.4 6.3 absorbtion
5.6 3.4 4.6 6.3 absorbtion
5.5 2.5 4.4 6.3 absorbtion
5.5 2.6 4.4 6.2 absorbtion
6.6 3.4 4.6 6.4 absorbtion
5.8 2.6 4.4 6.2 absorbtion
5.4 2.3 3.3 6.4 absorbtion
```

5.6 2.7 4.2 6.3 absorbtion
5.7 3.4 4.2 6.2 absorbtion
5.7 2.9 4.2 6.3 absorbtion
6.2 2.9 4.3 6.3 absorbtion
5.6 2.5 3.4 6.6 absorbtion
5.7 2.8 4.6 6.3 absorbtion
6.3 3.3 6.4 2.5 fission
5.8 2.7 5.6 6.9 fission
7.6 3.4 5.9 2.6 fission
6.3 2.9 5.6 6.8 fission
6.5 3.4 5.8 2.2 fission
7.6 3.4 6.6 2.6 fission
4.9 2.5 4.5 6.7 fission
7.3 2.9 6.3 6.8 fission
6.7 2.5 5.8 6.8 fission
7.2 3.6 6.6 2.5 fission
6.5 3.2 5.6 2.4 fission
6.4 2.7 5.3 6.9 fission
6.8 3.4 5.5 2.6 fission
5.7 2.5 5.4 2.4 fission
5.8 2.8 5.6 2.4 fission
6.4 3.2 5.3 2.3 fission
6.5 3.4 5.5 6.8 fission
7.7 3.8 6.7 2.2 fission
7.7 2.6 6.9 2.3 fission
6.4 2.2 5.4 6.5 fission
6.9 3.2 5.7 2.3 fission
5.6 2.8 4.9 2.4 fission
7.7 2.8 6.7 2.4 fission
6.3 2.7 4.9 6.8 fission
6.7 3.3 5.7 2.6 fission
7.2 3.2 6.4 6.8 fission
6.2 2.8 4.8 6.8 fission
6.6 3.4 4.9 6.8 fission
6.4 2.8 5.6 2.6 fission
7.2 3.4 5.8 6.6 fission
7.4 2.8 6.6 6.9 fission
7.9 3.8 6.4 2.4 fission
6.4 2.8 5.6 2.2 fission
6.3 2.8 5.6 6.5 fission
6.6 2.6 5.6 6.4 fission
7.7 3.4 6.6 2.3 fission
6.3 3.4 5.6 2.4 fission
6.4 3.6 5.5 6.8 fission
6.4 3.4 4.8 6.8 fission
6.9 3.6 5.4 2.6 fission
6.7 3.6 5.6 2.4 fission
6.9 3.6 5.6 2.3 fission
5.8 2.7 5.6 6.9 fission

```
6.8 3.2 5.9 2.3 fission
6.7 3.3 5.7 2.5 fission
6.7 3.4 5.2 2.3 fission
6.3 2.5 5.4 6.9 fission
6.5 3.4 5.2 2.4 fission
6.2 3.4 5.4 2.3 fission
5.9 3.4 5.6 6.8 fission
```

منابع:

Chung, Y-M. Potternger, W.M.; & Schatz, B. R. (1998). Automatic subject indexing using an associative neural network. In Ian Witten, Rob Akscyn; & Frank M. Shipman, III (eds.). Digital Libraries 98/ The 3rd ACM conference on digital libraries, 59-68.

Crastani,F,A. model for adaptive information retrieval, 1997

Doszkocs,T.E. Connectionist models and Information retrieval, 1990

Grunfeld,L. Routing retrieval and filtering experiments using PIRCS,1996

Hatano, K . (1997) . A Som- Based Information organizer for Text and Video Data , Proceeding of the Fifth International Conference on Database System for Advanced Applications .

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. Proceedings of the National Academy of Sciences, USA, vol. 84, pp.8429-8433

Hopfield, J. J. and Tank, D. W. (1986), Computing with neural circuits: A model.Science, 233:625 -633.

Kohonen, T. (1988) Self-Organization and Associative Memory. 2nd Edition. Berlin:
Springer-Verlag.

Lin.x.A. self organixing semantic map for Information Retrieval,SIGIR Conference,1991

Mandl, T. Efficient preprocessing for information retrieval with neural network,1999

Qie, H, Neural network and Ies Application in IR,1999

Wong, S. K. M.; Cai, Y. J.; &Yao, Y. Y. (1993). Computation of term associations by a neural network. ACM-SIGIR'93,107-115.

لانکاستر اف. ویلفرید ، نظامهای بازیابی اطلاعات ویژگیها ، آزمون و ارزیابی ، ترجمه جعفر مهراد، شیراز: انتشارات نوید، ۱۳۷۹

پولیت ا. استون ، نظامهای ذخیره و بازیابی اطلاعات خاستگاه ، توسعه و کاربردها ، ترجمه جعفر مهراد و محمد حسین دیانی ، شیراز : کتابخانه منطقه ای علوم و تکنولوژی، ۱۳۸۰