پژوهشگاه علوم انسانی و مطالعات فرهنگی

انجمن زبان شناسی ایران

هم اندیشی زبان فارسی در رایاسپهر

استخراج هستان شناسی برای دانشنامه معماری و شهرسازی

یحیی تابش

دانشگاه صنعتی شریف

حامد ملک

دانشگاه صنعتی امیرکبیر

۱۰ دی ماه ۱۳۸۸

معرفی پروژه

- تشکیل پایگاه معرفتی عظیمی برای تاریخ معماری ایران از هزاره هفتم پیش از میلاد تا کنون
- انتشار محصول این طرح به صورت وبگاه
- دامنه جغرافیایی طرح همه جهان ایرانی (از سیحون تا خلیج فارس و از سند تا فرات) است.
- دامنه تاریخی موضوع طرح از روزگار پیش از تاریخ تا پایان دوره پهلوی اول را شامل می شود.

معرفی پروژه

- دانشنامه متعارف ابزاری است که پیشینه تحقیق درباره عنوانی معین به نام «مدخل» را در اختیار عموم یا پژوهشگران قرار می دهد.
- در این دانشنامه همه محققان مؤلف بالقوه مقاله درباره هر مدخل شمرده می شوند.
- در این دانشنامه مداخل ثابت نیست و به ازای جستجو و نیاز مخاطبان، پیوسته بر مداخل افزوده می شود.
- اساس کار در این دانشنامه امکان جستجوی دقیق در همه مدارک مرتبط با تاریخ معماری و شهرسازی ایران زمین است.
- مجموعه اطلاعات به دو دسته «مدارک» و «مداخل» تقسیم می شوند.
- مدخل: واژه ای که اطلاعت حول آن گرداوری یا تولید می شود.
  مدرک: هر نوع منبع چاپی یا غیر چاپی مانند کتاب، فیلم، نسخه خطی و غیره است.

معرفی پروژه

- مجری این طرح فرهنگستان هنر و حامی اصلی آن وزارت مسکن و شهرسازی است.
- همکاران طرح: سازمان میراث فرهنگی، سازمان اسناد و کتابخانه ملی و دانشگاه شهید بهشتی
- طرح اکنون در مرحله تأسیس قرار دارد که هدف آن تهیه ۵۰۰۰ مدخل اولیه در وبگاه است.
- در مرحله کنونی زبان دانشنامه فارسی و انگلیسی است که در آینده به همه زبان هایی که اطلاعاتی در آن زبان وجود دارد گسترش خواهد یافت.

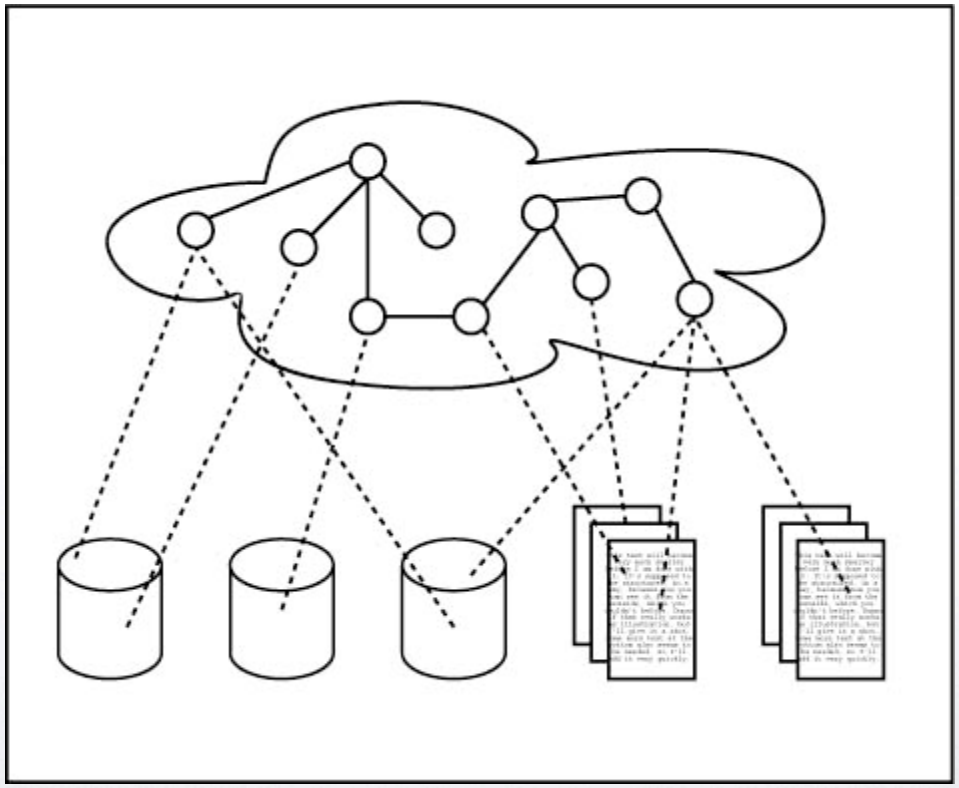
معرفی پروژه

http://www.eiah.org

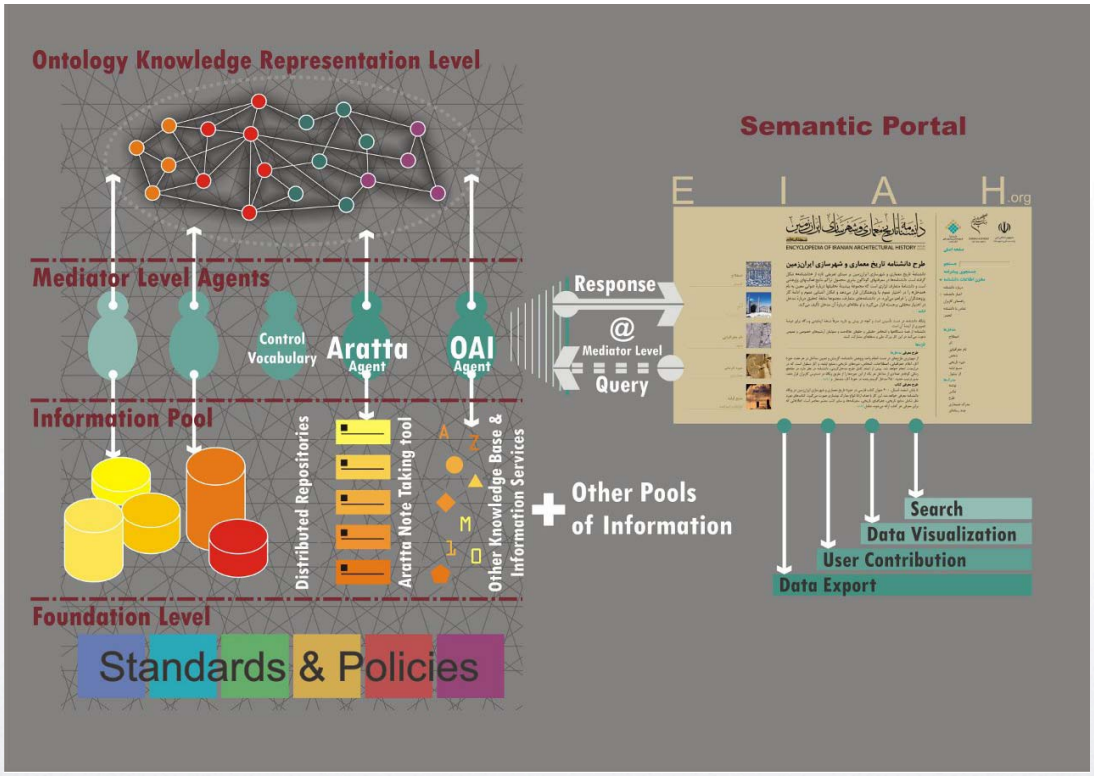
معماری اطلاعاتی

- استاندارد مورد استفاده برای بازنمایی و تبادل دانش استاندارد «مدل موضوعی» یا Topic Maps است.
- در این مدل، اطلاعات با استفاده از سه مفهوم موضوعات، وابستگی ها و رخدادها بازنمایی می شوند.
- موضوع: نماینده یک مفهوم مانند شخص، مکان، رویداد و غیره
  وابستگی: نحوه ارتباط بین موضوعات
  رخداد: نمایش رابطه بین موضوعات و منابع اطلاعاتی مرتبط
- مدل موضوعی را می توان نوع خاصی از فناوری وب معنایی دانست.

معماری اطلاعاتی



معماری اطلاعاتی



هستان شناسی

هستان شناسی سبک

هستان شناسی سنگین

مدل های مفهومی

(تعریف طبقات، ویژگی و روابط آنها)

معماری اطلاعاتی

پروژه اَرَته

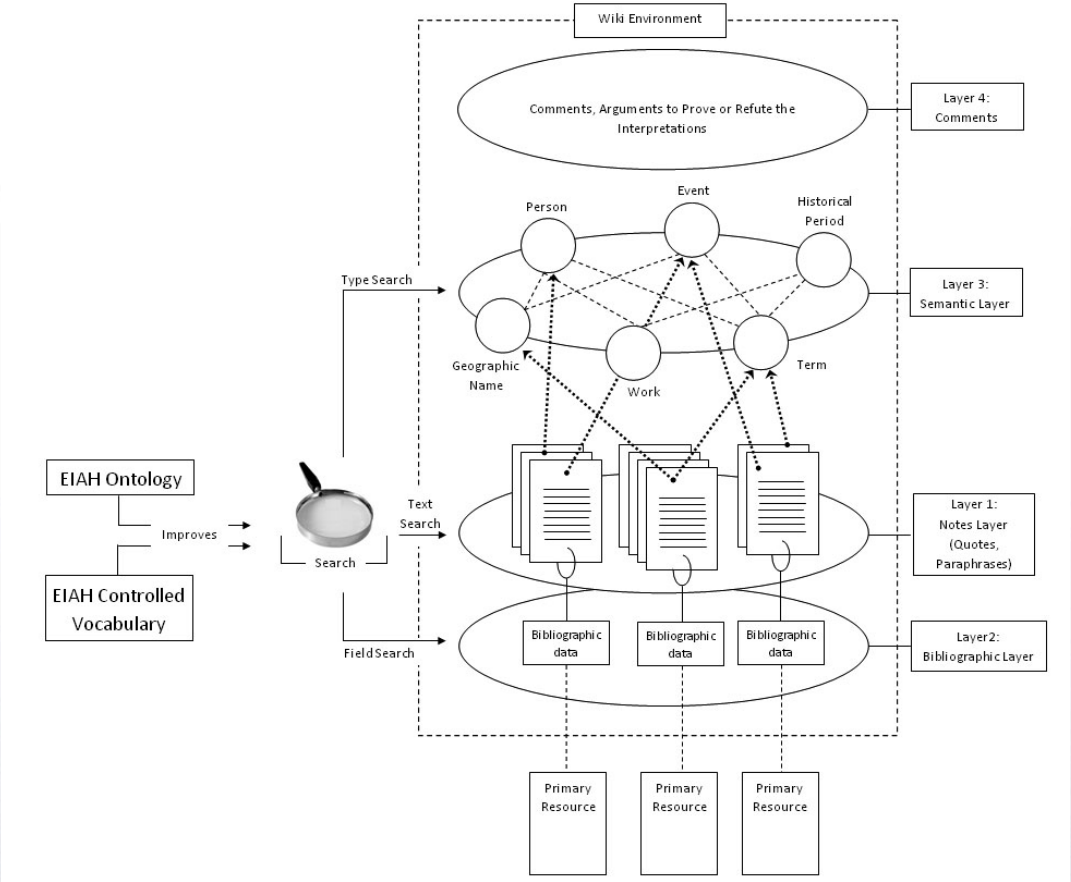
جستجوی تقریبی

پروژه استخراج اسامی خاص

پروژه اَرَته

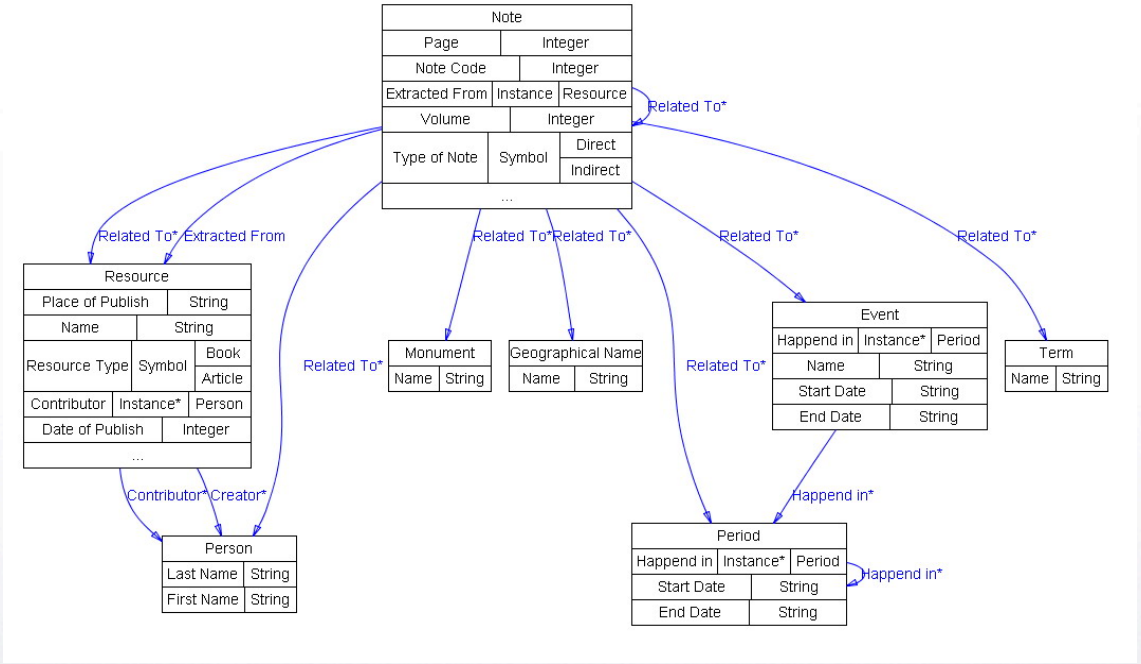
- هدف: ایجاد یک کتابخانه معنایی از یادداشت های تحقیقاتی
- نام پروژه برگرفته از نام تمدن باستانی گمشده ای به نام اَرَته است.
- بیش از ۱۰ هزار یادداشت تحقیقاتی

پروژه آرَته



نمودار مدل داده

پروژه اَرَته



هستان شناسی اَرَته

پروژه اَرَته

- مدیاویکی (Mediawiki)
- سمنتیک مدیاویکی (Semantic Mediawiki)
- سمنتیک فرم (Semantic Form)
- استاندارد دابلین کور (Dublin Core)

جستجوی تقریبی

- هدف: یافتن عبارات و زیررشته هایی که تفاوت قابل اغماض با رشته مورد جستجو دارند.
- جستجوی تقریبی بی نقص نیاز وسیع به اطلاعات زبانی، واژگانی، معنایی و مانند آن دارد.
- در این پروژه به پیاده سازی شیوه هایی که توسط الگوریتم ها و جداول ساده قابل پیاده سازی هستند پرداخته شده است.
- برای این منظور از توصیف «جستجوی تقریبی برای زبان فارسی ایران» تهیه شده توسط شرکت فارسی وب شریف استفاده شده است.

جستجوی تقریبی

- غلط های کدگذاری: مانند استفاده از نااستاندارد از حروف عربی به جای حروف فارسی
- غلط های املایی معمول یا نامحسوس: مانند نوشتن کلمه «مؤمن» به شکل «مومن»
- تفاوت در اِعراب گذاری عبارات و وجود یا عدم وجود نویسه های ضمنی
- همه موارد فوق در قالب این پروژه به موتور جستجوی آزاد لوسن «Lucene» پیاده سازی شده و قابل استفاده برای عموم است.

تشخیص اسامی خاص

- با توجه به حجم بالای متون نوشتاری و نیاز به برچسب گذاری آنها برای تولید مناسب هستان شناسی، استفاده از روش های خودکار یک الزامی است.
- ترجیح روش های آماری بر روش های زبان شناسی به دلیل استقلال روش از زبان
- انتخاب کتابخانه معنای اَرَته به عنوان منبع اصلی استخراج اسامی خاص
- پیاده سازی الگوریتم با کمک ابزارهای آزاد موجود

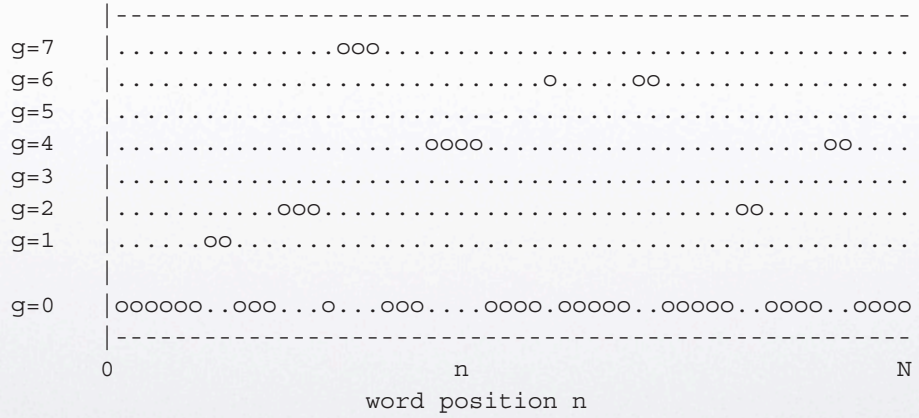
تشخیص اسامی خاص

- اسامی به ۶ دسته یا کلاس تقسیم می شوند.
- اصطلاح: آتشگاه، خاتم کاری، کاروان سرا
- اثر: معبد آناهیتا، مسجد جامع اصفهان، حمام علی قلی آقا
- نام جغرافیایی: آباده، دریای مازندران، کرمان
- دوره تاریخی: آل بویه، ساسانیان، صفویه
- شخص: سلطان سنجر، لطف علی خان زند، شاه عباس
- منبع اولیه: آثار عجم، سلجوق نامه، سفرنامه ناصر خسرو

تشخیص اسامی خاص

دنباله ورودی: $w_1^N = w_1 \dots w_n \dots w_N$

دنباله خروجی: $g_1^N = g_1 \dots g_n \dots g_N$



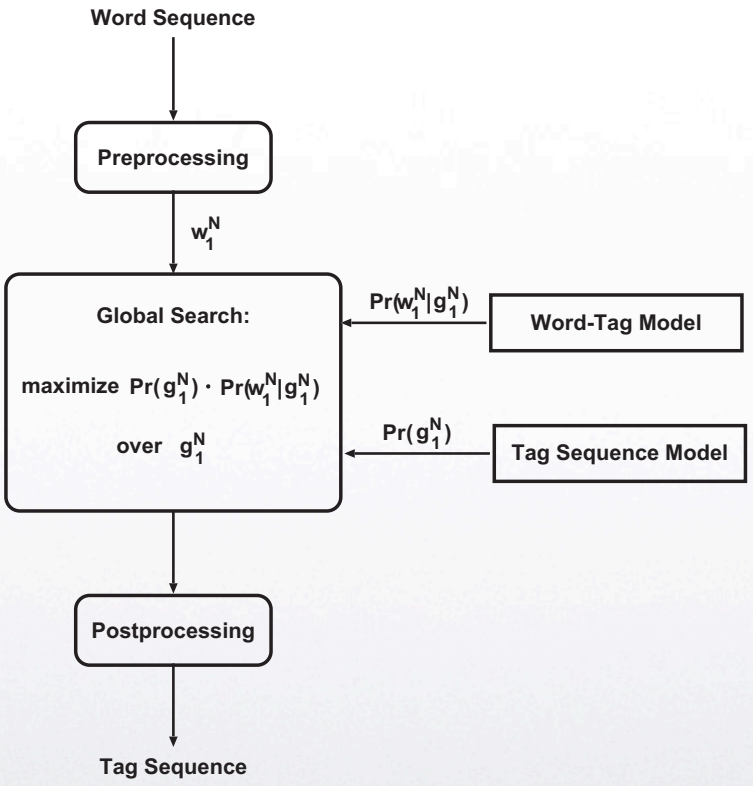
تشخیص اسامی خاص

ایده اصلی مشابه روش معمول آماری برای POS Tagging است.

قانون تصمیم گیری بیز:

$$
\begin{aligned}
w_1^N \rightarrow \hat{g}_1^N &= \arg\max_{g_1^N} Pr(g_1^N | w_1^N) \\
&= \arg\max_{g_1^N} \{Pr(g_1^N, w_1^N)\} \\
&= \arg\max_{g_1^N} \{Pr(g_1^N) \cdot Pr(w_1^N | g_1^N)\}
\end{aligned}
$$

تشخیص اسامی خاص



تشخیص اسامی خاص

$$\arg\max_{g_1^N} Pr(g_1^N|w_1^N) = \arg\max_{g_1^N} \{Pr(g_1^N) \cdot Pr(w_1^N|g_1^N)\}$$

احتمال تعلق: مدل مرتبه صفر

$$Pr(w_1^N|g_1^N) = \prod_{n=1}^{N} p_0(w_n|g_n)$$

احتمال بایگرام POS: مدل مرتبه یک

$$Pr(g_1^N) = \prod_{n=1}^{N} p_1(g_n|g_{n-1})$$

آموزش: هر دو توزیع احتمالاتی بالا را با کمک نمونه های دستی تخمین می زنیم.

تشخیص اسامی خاص

$$p_0(w|g) = \frac{p(w,g)}{p(g)} = \frac{p(w)\cdot p_0(g|w)}{p(g)}$$

مسأله بهینه سازی نهایی:

$$\arg\max_{g_1^N} Pr(g_1^N|w_1^N) = \arg\max_{g_1^N} \left\{ \prod_{n=1}^{N} \left[\frac{p_0(g_n|w_n)}{p(g_n)} \cdot p_1(g_n|g_{n-1}) \right] \right\}$$

برای حل مسأله بهینه سازی بدست آمده از روش برنامه ریزی پویا استفاده می کنیم.

تشخیص اسامی خاص

- تفاوت اصلی با POS Tagging در طول برچسب ها است که می تواند بیش از یک باشد.
- برای این منظور از یک مدل ساده طول استفاده می شود:
- هر برچسب با دو برچسب «شروع» یا «ادامه» جایگزین می شود.

تشخیص اسامی خاص

قانون تصمیم گیری بیز:

$$
\begin{aligned}
w_1^N \rightarrow \hat{g}_1^N &= \arg\max_{g_1^N} Pr(g_1^N|w_1^N) \\
&= \arg\max_{g_1^N} \left\{ Pr(w_1^N) \cdot Pr(g_1^N|w_1^N) \right\} \\
&= \arg\max_{g_1^N} \left\{ Pr(g_1^N, w_1^N) \right\}
\end{aligned}
$$

$$
\begin{aligned}
Pr(g_1^N, w_1^N) &= \prod_n Pr(g_n, w_n|g_1^{n-1}, w_1^{n-1}) \\
&= \prod_n \left[Pr(g_n|g_1^{n-1}, w_1^{n-1}) \cdot Pr(w_n|g_1^n, w_1^{n-1}) \right]
\end{aligned}
$$

$$
\begin{aligned}
Pr(g_1^N, w_1^N) &= \prod_n Pr(g_n, w_n|g_1^{n-1}, w_1^{n-1}) \\
&= \prod_n \left[Pr(g_n|g_1^{n-1}, w_1^{n-1}) \cdot Pr(w_n|g_1^n, w_1^{n-1}) \right]
\end{aligned}
$$

تشخیص اسامی خاص

قانون تصمیم گیری بیز نهایی:

$$\arg\max_{g_1^N} Pr(g_1^N|w_1^N) = \arg\max_{g_1^N} \left\{ \prod_n \left[p(g_n|g_{n-1}, w_{n-1}) \cdot p(w_n|g_{n-1}^n, w_{n-1}) \right] \right\}$$

فرض اضافه:

$$p(w_n|g_{n-1}^n, w_{n-1}) = \begin{cases} p(w_n|g_n) & g_n = g_{n-1} \\ p(w_n|g_n, w_{n-1}) & g_n \neq g_{n-1} \end{cases}$$

تشخیص اسامی خاص

تعداد فیش های آموزش	500
تعداد کلمات آموزش	56343
تعداد فیش های آزمایش	100
تعداد کلمات آزمایش	12330

اصطلاح	2862
اثر	1347
نام جغرافیایی	842
دوره تاریخی	107
شخص	392
منبع اولیه	101

کارهای آینده

- هوشمندسازی الگوریتم های جستجو با کمک روش های پردازش زبان طبیعی
- استفاده از تکنیک های آماری بهبود کیفیت تشخیص اسامی خاص
- استفاده از تکنیک های زبان شناسی در الگوریتم تشخیص اسامی خاص

اطلاعات بیشتر

- وبگاه دانشنامه معماری و شهرسازی ایران زمین: www.eiah.org
- وبگاه پروژه اَرَته: http://sohrab.eiah.org:3080/aratta

hmalek@aut.ac.ir

tabesh@sharif.ir