

How does a site like kayak.com aggregate content? [closed]

Asked 13 years, 11 months ago Modified 7 years, 2 months ago

Viewed 80k times



84



Closed. This question needs to be more [focused](#). It is not currently accepting answers.



Want to improve this question? Update the question so it focuses on one problem only by [editing this post](#).

Closed 8 years ago.

[Improve this question](#)

Greetings, I've been toying with an idea for a new project and was wondering if anyone has any idea on how a service like Kayak.com is able to aggregate data from so many sources so quickly and accurately. More specifically, do you think Kayak.com is interacting with APIs or are they crawling/scraping airline and hotel websites in order to fulfill user requests? I know there isn't one right answer for this sort of thing but I'm curious to know what others think would be a good way to go about this. If it helps, pretend you are going to create kayak.com tomorrow ... where is your data coming from?

[api](#)[architecture](#)[screen-scraping](#)[aggregate](#)[Share](#)[Improve this question](#)[Follow](#)

asked Jan 5, 2011 at 17:27



Jeff

2,828 ● 3 ● 29 ● 31

7 Answers

Sorted by:

Highest score (default)



149



I'm working in travel industry as a software architect / project lead on the precisely kind of project you describe - in our region we work with suppliers directly, but for outgoing we connect to several aggregators.

To answer your question... some data you have, some you get in various ways, and some you have to torture and twist until it confesses.

What's your angle?

The questions you have to ask are... Do you want to sell advertising like Kayak or do you take a cut like Expedia? Are you into search or into selling travel services? Do you target niche (for example, just air travel) or everything (accommodation, airlines, rent-a-car, additional services like transport/sightseeing/conferences etc)? Do you target region (US or part of US) or the world? How deep do you go - do you just show several sites on a single screen, or

do you bundle different services together and package them dynamically?

Getting the data

If you're going with Kayak business model, you technically don't need site's permission... but a lot of sites have affiliate programs with IFrames or other simple ways to direct the customer to their site. On the plus side, you don't have to deal with payments/complaints and travelers themselves. As for the cons... if you want to compare prices yourself and present the cheapest option to the user, you'll have to integrate on a deeper level, and that means APIs and web scraping.

As for web scraping... avoid it. It sucks. Really. Just don't do it. Trust me on this one. For example, some things like lowcosters you can't get without web scraping. Low cost airlines live from value added services. If the user doesn't see their website, they don't sell extra stuff, and they don't earn anything. Therefore, they don't have affiliates, they don't offer APIs, and they change their site layout almost constantly. However, there are companies which earn a living by web scraping lowcoster's sites and wrapping them into nice APIs. If you can afford them, you can give your users cost-comparison of low cost flights and that's huge.

On the other hand, there are "normal" carriers which offer APIs. It's not that big of a problem to get to airlines since they're all united under [IATA](#); basically, you buy from

IATA, and IATA distributes the money to carriers.

However, you probably don't want to connect directly to carrier network. They have web services and SOAP these days, but believe me when I say that there are SOAP protocols which are just an insanely thin wrappers around a text prompt through which you can interact with a mainframe with an 80es-style protocol (think of a Unix prompt where you're billed per command; and it takes about 20 commands to do one search). That's why you probably want to connect to somebody a bit more down the food chain, with a better API.

Airlines are thus on both extremes of Gaussian curve; on one side are individual suppliers, and on the other highly centralized systems where you implement one API and you're able to fly anywhere in the world. Accommodation and the rest of travel products are in between. There are several big players which aggregate hotels, and a ton of small suppliers with a lot of aggregators which cover only part of a spectrum. For example, you can rent a lighthouse and it's even not that expensive - but you won't be able to compare the prices of different lighthouses in one place.

If you're into Kayak business model, you'll probably end up scraping websites. If you're into integrating different providers, you'll often work with APIs, some of which are pretty good, and most of which are tolerable. I haven't worked with RSS but there's not a lot of difference between RSS and web scraping. There is also a fourth option not mentioned in Jeff's answer... the one where

you get your data nightly, for example .CSV files through FTP and similar.

Life sucks (mini-rant)

And then there's complexity. The more value you want to add, the more complexity you'll have to handle. Can you search accommodations which allow pets? For a hostel which is located less than 5 km from the town center? Are you combining flights, and are you able to guarantee that the traveler will have enough time to get from one airport to another... can you sell the transport in advance? A famous cellist doesn't want to part from his precious 18th century cello; can you sell him another seat for the cello (yep, not making this one up)?

Want to compare prices? Sure, the room is EUR 30 per night. But you can either get one double for 30 and one single for 20, or you can get one extra bed in a double and get 70% off for third person. But only if it's a child under 12 years of age; our extra beds are not for adults. And you don't get the price for extra bed in search results - only when you calculate the final price.

And don't even get me started on dynamic packaging. Want to sell accommodation + rent-a-car? No problem; integrate with two different providers, and off you go... manually updating list of locations in the city (from rent-a-car provider) to match with hotels (from accommodation provider, who gives you only the city for each hotel). Of course, provided that you've already matched the list of

cities from the two, since there is no international standard for city codes.

Unlike a lot of other industries which have many products, travel industry has many very complex products. Amazon has it easy; selling books and selling potatoes, it's the same thing; you can even ship them in the same box. They combine easily and aren't assembled from many parts. :)

P.S. Linking to an interesting recent thread on Hacker News with some [insider info regarding flights](#). P.P.S. Recently stumbled on a great albeit rather old blogpost on [IATA's NDC protocol with overview of how travel industry is connected and a history lesson how this came to be](#).

Share Improve this answer

edited Oct 5, 2017 at 12:03

Follow

answered Jan 6, 2011 at 6:02



Domchi

10.8k ● 7 ● 56 ● 64

Domchi, has this changed a lot this year? Are there other APIs available now? – [Rizwan Kassim](#) Dec 28, 2011 at 21:46

No, not a lot; the market is fragmented and this is unlikely to change soon, if ever. Notable event is Google entering the flights market (through ITA Software, see mavrcks answer); they have the resources to consolidate the market and offer APIs but I doubt that's what they intend/are able to do. In the startup world, most interesting contender is probably

airbnb.com but so far they don't offer API. APIs in this domain are not hard to find, but are rarely free. Check programmableweb.com/apitag/booking and programmableweb.com/apitag/travel for a good API list.
– Domchi Jan 1, 2012 at 16:39

Are you saying that Kayak.com scrapes airlines' sites for content, and doesn't have to pay for it? What about their business model allows them to do that? The terms of use for the major airlines seem to say that one cannot scrape content/data from their site for use on another site.
– Ryan Bales Nov 26, 2012 at 23:03

@Ryan I can't say what Kayak does, but I know that most low-cost airlines don't offer any APIs in order to drive sales exclusively through their websites and upsell as much as they can. So in turn, aggregators scrap their websites and simulate user interaction. In response airlines frequently change website structure and the game of cat and mouse goes on. I would imagine that it's pretty hard for carriers to prove that this goes on, but they know about it and probably don't want to prevent it altogether since they do want the traffic they otherwise wouldn't get. – Domchi Nov 29, 2012 at 8:58

@Domchi how is what you're describing legal? From what I've read, web scraping has been deemed illegal in various court cases regarding airline data, ebay listings, and others (especially for commercial use of said data). – Justin Skiles Feb 26, 2015 at 14:28 ✎



9

They use a software package like [ITA Software](#), which is one of the companies Google is in the process of picking up.



Follow



JSW189

6,315 ● 11 ● 46 ● 73



answered Jan 6, 2011 at 3:23



mavrck

1,923 ● 1 ● 10 ● 15

-
- 1 scooped up in the meantime, see developers.google.com/gpx-express/v1/trips/search for an API – [wires](#) Aug 6, 2014 at 17:27
-



Only 3 ways I know of to get data from websites.

8



RSS Feeds - We use rss feeds a lot at my company to integrate existing site's data with our apps. It's fast and most sites already have an RSS feed available. The problem with this is not all sites implement the RSS standard properly so if you're pulling data from many RSS feeds across many sites then make sure you write your code so that you can add exceptions and filters easily.



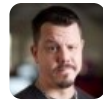
APIs - These are nice if they are designed well and have all the information you need, however that's not always the case, plus if the sites are not using a standard api format then you'll have to support multiple API's.

Web Scraping - This method would be the most unreliable as well as the most expensive to maintain. But if you're left with nothing else it can be done.

Share Improve this answer

answered Jan 5, 2011 at 21:58

Follow



Jeff Busby

2,011 ● 17 ● 12



3



[This article](#) says that Kayak was asked to stop scrapping a certain airlines page. That leads me to believe that they probably do scraping on sites that they don't have a relationship with (and a data feed that comes with that relationship).



Share Improve this answer

answered Apr 13, 2012 at 21:52



Follow



Jake Wilson

91k ● 96 ● 259 ● 371



3



Travelport offer a product called "Universal API" which connects to flights and hotels and car rental companies and copes with package deals and all the various complexities to do with taxes and exchange rates:

<https://developer.travelport.com/app/developer-network/resource-centre-uapi>

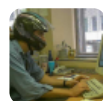


I've just started using it and it seems fine so far. The queries are a little slow, but then so is every query on every OTA (Online travel agent)'s site.

Share Improve this answer

answered Nov 3, 2013 at 9:02

Follow



Tim Cooper


10.5k ● 6 ● 65 ● 78

Whats the cost of using Universal API? – [Amit](#) Jan 6, 2016 at 21:32

- 1 I've forgotten. A one-off fee in the low thousands plus an annual fee in the low thousands, I think. – [Tim Cooper](#) Jan 8, 2016 at 3:55
-

Is universal API available for mobile? – [iSrinivasan27](#) Jan 25, 2016 at 5:43

uAPI utilizes SOAP protocol. There is no restrictions per IP address range but I wouldn't integrate API credentials into publicly available mobile apps. Build your own middleware to protect credentials.

support.travelport.com/webhelp/uapi/uapi.htm – [gavenkoa](#) Jan 6, 2021 at 11:16 



There's two good APIs I've found from flight comparison websites recently

2



There's one from [Wego](#), and one from [Skyscanner](#). Both seem to have a good range and breadth of data from a number of airlines and good documentation too.



Wego pays each time a user clicks from your app to a booking website and Skyscanner pay affiliates 50% of 'revenue' (I assume that means the commission they make from airlines)

Share Improve this answer

answered May 27, 2014 at 21:23

Follow



[Jonathon Blok](#)

749 ● 4 ● 15

4 FYI Wego also charge \$1000 USD per year for the privilege of using their API. – [Sk446](#) Jun 3, 2014 at 18:14 ✎

Correct. Also, further to my post (as I know a bit more now), Skyscanner pay per exit click in the same way Wego do.
– [Jonathon Blok](#) Jun 4, 2014 at 10:16

2 Hello, Skyscanner API team here. Hotels is in the works. Please visit business.skyscanner.net if you'd like to discuss using any of our APIs (flights, car hire, hotels). – [Skyscanner](#) Jun 11, 2014 at 16:28

1 Just checked SkyScanner. New devs are worthless to them. You can't use their Flights API unless your site generates at least 200,000 unique visits a month! Pathetic! – [Hajjat](#) Dec 16, 2015 at 16:00

1 Its been 4 years and @Skyscanner still doesn't provide hotel API. – [nad](#) Jul 29, 2018 at 22:50



1



This is an old post but I thought I'd just add. I'm a data architect who works for a company that feeds these travel sites with content. This company enters into contracts with many hotel brands, individual hotels and other content providers. We aggregate this information then pass it onto the different channels. They then aggregate again in to their system. The Large GDS systems are also content providers. Aggregation is done by many methods... matching algorithms(in-house) and keys. Being an aggregation service, we need to communicate on the client level.

Hope this helps! cheers!

Share Improve this answer

answered Aug 16, 2016 at 13:10

Follow



Krdls

31 ● 1
